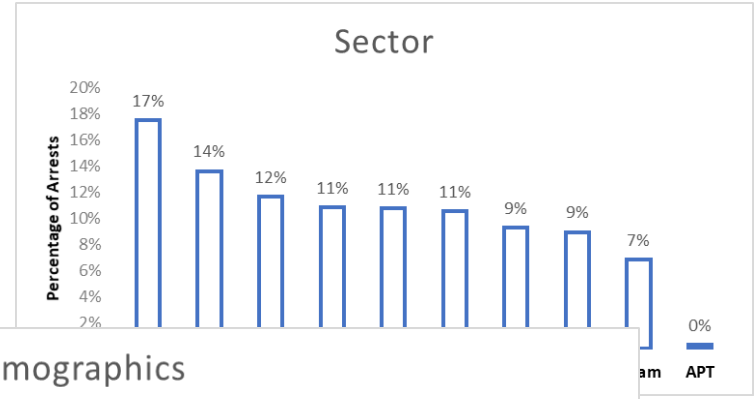
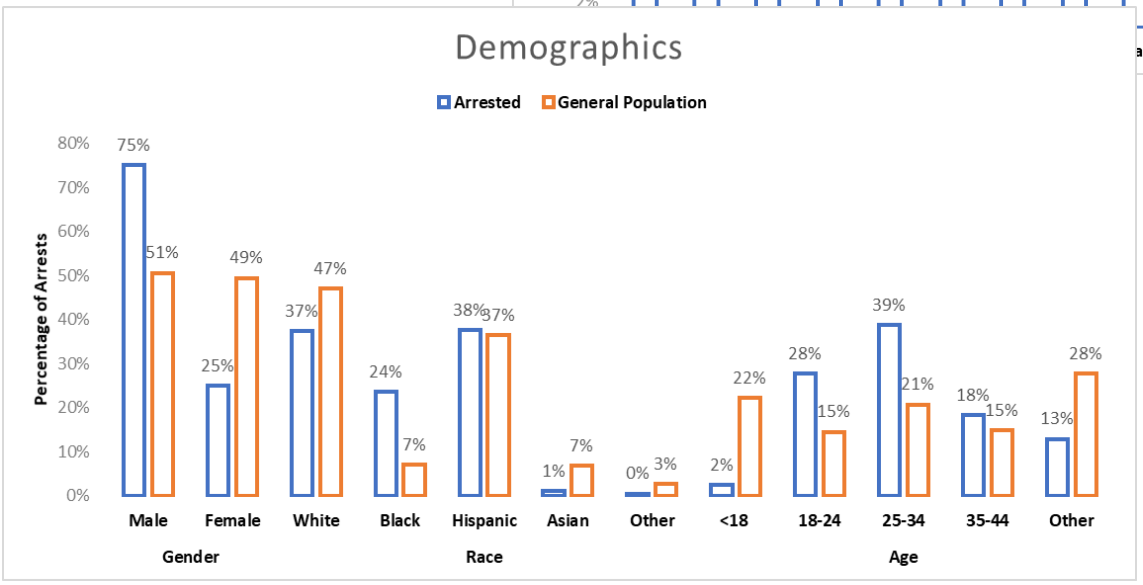
METHODS

APD released supporting documentation alongside its Annual Racial Profiling Report for the first time in 2016. In the spirit of transparency, citizens can access comprehensive data that supports yearly departmental summaries. Available information includes demographics along with temporal and geographic identifiers. Through regression analytics, I find that APD differs from other major cities in the way it targets minorities. Most cities focus police activity on areas of criminal activity, such as housing projects and poor neighborhoods. In Austin, location is irrelevant. Instead, minorities are targeted according to time-of-day policing. Blacks and Hispanics are far more likely than whites to be arrested in the night hours, even after controlling for criminal activity.

Data

Of all arrests, 9,126 contain full demographic data. Demographic characteristics are (1) *gender*, (2) *age,* and (3) *race*. Regarding *gender*, 6,847 (75.03%) are male, while 2,279 (24.97%) are female. Males are nearly three times as likely to be arrested as females. Regarding age, the three most likely groups to be arrested are 18-24 (2,523, 27.65%), 25-34 (3,535, 38.74%), and 35-44 (1,671, 18.31%). The highest likelihood of arrest occurs between the ages of 18 and 44. APD arrested 220 juveniles under the age of 18, making up 2.41% of those incarcerated. Juvenile offenders are 17.5 times less likely to be arrested than personnel in the 18-24 age group. Regarding race, three groups



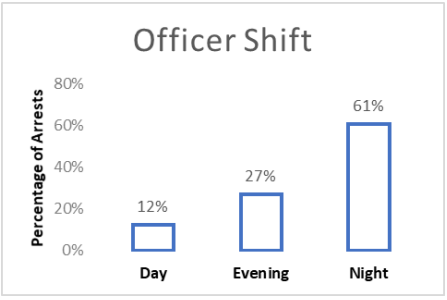


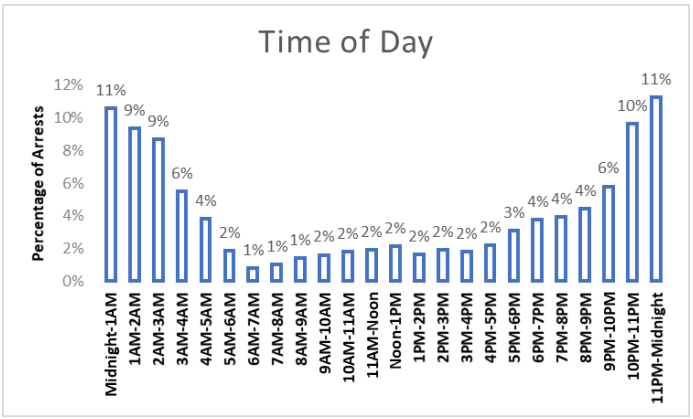
compose the majority of arrests: Hispanic (3,430, 37.58%), white (3,412, 37.39%), and Black (2,146, 23.52%). Other arrested racial groups include Asian (93, 1.02%), Middle Eastern (35, 0.38%), Hawaiian/Pacific Islander (5, 0.05%), and American Indian/Alaskan Native (5, 0.05%). Race is identified by the officer at arrest then confirmed by the suspect at booking. Hispanics and Blacks are arrested at rates 1.3 and 4.2 times higher than whites respectively. Whites are the third most likely race to be arrested.

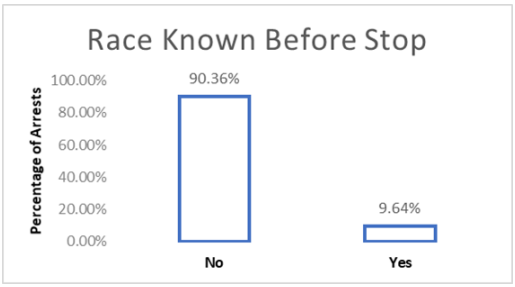
Of all arrests, 9,122 contain full geographic data. APD divides Austin into 10 sectors, each sector identified by a code name. Code names correspond approximately to an established neighborhood. Most arrests occurring in Edward (Rundberg, 17.35%) and Frank (Del Valle,

13.50%). Following in decreasing proportion of arrests are Charlie (East Austin, 11.53%), Henry (Montopolis, 10.71%), David (Barton Springs, 10.68%), and Baker (West Campus, 10.47%). Neighborhoods with highest the highest arrest rates are also predominately minority inhabited.

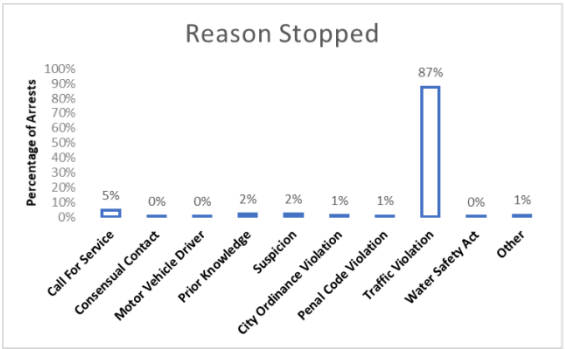
All 9,184 arrests contain full temporal data. Temporal factors are (1) date and (2) time. Date can be divided into month and season. Time can be divided into hours and officer shifts. Regarding date, most arrests occur in January (11.56%), while the least occur in November (7.21%). Most arrests occur in the winter (28.92%) while the least occur in the spring (22.80%). Regarding time, most arrests occur from 11PM to 12AM (11.24%). The least take place from 6AM to 7AM (0.78%). Most arrests take place in the night time hours, with fewest occurring in hours of sunshine. Organizationally, APD breaks officer patrolling into three time-shifts: Day (6AM to 2PM), evening (2PM to 10PM) and night (10PM to 6AM). 60.63% of arrests take place during the night shift but only 12.40% in the day shift.

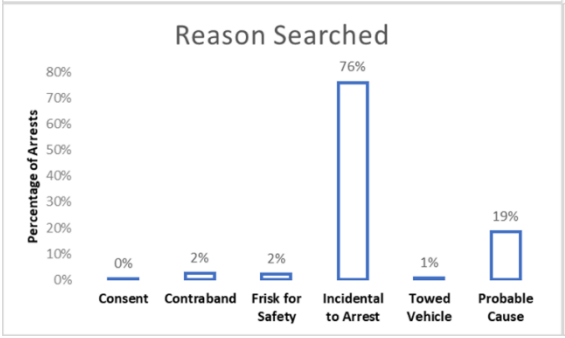


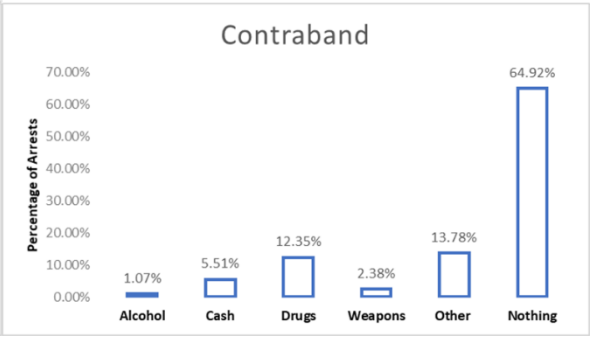




Of all arrests, 7,280 contain full criminal data. Criminal data categories are *Reason Stopped*, *Reason Searched*, *Contraband Found During Search*, and *Race Known Before Stop*. Regarding *reason stopped*, the most common is violation of transportation code or vehicle laws (87.47%). Surprisingly, violation of penal codes and suspicious drivers only account for 0.77% and 2.05%. Regarding *Reason Searched*, searches that are incidental to arrest account for 76.03% of searches. The most common non-procedural reason to be searched is probable cause (18.56%). The least likely reason is consent, indicating that most people arrested are not complicit in the search. C*ontraband Found During Search* includes drugs (12.35%), cash (5.51%), and weapons (2.38%). However, most searches (64.92%) do not result in the discovery of contraband. Regarding *Race Known Before Stop*, in 90.36% race is not known, compared to 9.64% in which race is known. This is measured in two ways. First, dispatchers may send racial information to the officer in the event of a call for service. Second, an officer annotates in an after-action report whether they were aware of an arrested suspect's race. However, because most arrests result from routine traffic stops, implicit bias probably increases the likelihood that officers used racial categorization to conduct traffic stops. Probably, there is a much larger percentage of arrests that result from an officer knowing the suspect's race before the stop.

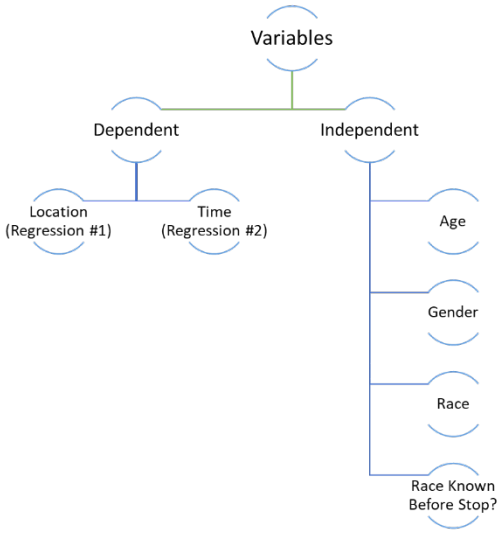






Analysis

To determine the relationship between demographic variables and arrest outcomes, I conducted multivariate linear regression analysis in Microsoft Excel. The original data contains continuous, binary, and categorical variables. All variables used in regressions were coded appropriately. I coded dependent variables as continuous and independent variables as either continuous or binary, with dummy variables applied to categorical independent variables that contain multiple categories of interest. Each variable is filtered to remove blanks. As a result, 6,054 arrests contain full demographic, geographic, temporal, and criminal data.



Racism exacerbates imbalanced arrests of minorities. Police officers are more likely to arrest Blacks and Hispanics regardless of criminal activity. By flooding Black and Hispanic parts of the city with patrollers, racial biases are pronounced and encouraged. Likewise, by concentrating officers during the night shift, officers are more likely to arrest individuals based on demographics.

Organizational factors upon which I conduct linear regression analysis are *Location* and *Time*. By understanding what factors influence location and time of arrest I can make policy recommendations that mitigate racism and subjectivity. For example, most police departments utilize hotspot policing to focus officers in locations with high likelihoods of criminal activity. In Austin, hotspot policing suggests that policing efforts should focus on sectors Edward and Frank where most criminal activity occurs. Using the same rationale, hotspots can also be temporal. Most arrests occur between 10PM nd 2AM. Austin police are therefore especially vigilant in the late-night hours. However, most arrests occur because of traffic violations. Possibly, police activity is misdirected, focusing on misdemeanor activity instead of major crimes.

Arrests are not indicative of criminal activity. It is faulty to assume that sectors Edward and Frank contain the most crime, when they may simply be targets of police activity. Likewise, it is faulty to assume that most crime occurs during the night hours, when in fact there may be more officers patrolling during the night hours and subsequently making arrests. Most arrests

occur from violation of transportation and vehicle laws, which is not traditionally a criminal activity. Calls for service indicate criminal activity, yet make up only 4.67% of arrests.

Therefore, my statistical analysis will indicate that criminal activity is not necessarily aligned with locations and times on which APD focuses. Rather, hotspot policing distracts from crime-fighting. If statistical analysis shows that police officers disproportionately arrest minorities in specific parts of town or at particular times of day, there is reason to believe that their efforts are misdirected. Minorities do not commit crimes at the same rate in which they are imprisoned. Their imprisonment should reflect their crimes. Police should focus on crime prevention, not the incarceration of minorities.

In the first regression, *Location* is the dependent variable. The independent variables are *Age*, *Race*, *Gender*, and *Race Known Before Stop*. *Location* is coded using weighted probability of occurrence per sector. There are 12 sectors. Two sectors were eliminated due to their negligible size. To code locational sectors, I first counted the total number of arrests that occurred in each sector. Then, for each sector, I divided arrests by total arrests. The resulting percentage is the probability of arrest in the given sector. For example, 1593 arrests occurred in sector Edward out of 9123 total arrests. The resulting probability is equal to 17.46%. I use respective probabilities for each sector, thus creating a continuous dependent variable.

In the second regression *Time* is the dependent variable. The independent variables are *Age*, *Race*, *Gender*, and *Race Known Before Stop*. *Time* is already a continuous variable. However, I converted each time block into a weighted probability that indicates likelihood of arrest. First, time is broken into 24 blocks, each block representing one hour. Second, number of arrests are tallied for each hour of the day, then divided by total arrests. For example, the block of time Midnight to 1AM contains 972 arrests. 972 divided by 9184 (total arrests) equals 10.58%. Similarly, probable percentages are attached to each arrest. Time is therefore a continuous dependent variable.

I considered combining time-of-day and month-of-year as a united temporal factor. January contains the highest number of arrests (1062, 11.56%) while April contains the least (627, 6.83%). There is significant monthly variation, but very little seasonal variation. The winter months contain 28.92% of all arrests while the least active season, spring, contains 22.80% of all arrests. I will not conduct regression using a combined temporal factor for two reasons. First, this variation is less interesting than the variation indicated according to time, due

to the disparity in arrests at specific hours of the day. Second, even though the combination of weighted percentages for both time and month offer an interesting perspective regarding demographic biases with relation to arrests, policies regarding time and month are separate entities. Policies dictating the time of day in which police officers more heavily police the streets are decided apart from the policies dictating the months or seasons in which police activity will be heaviest. For the sake of simplification in both statistical analysis and departmental policy recommendations, I chose not to create a combined temporal dependent variable.

Both regressions use the same independent variables: *Age*, *Race*, *Gender*, and *Race Known Before Stop*. The variable *Age* is itself already a continuous variable and does not need to be coded. The mean age is 31 years. For *Race*, the raw data contains 8 categories: American Indian/Alaskan Native, Asian, Black, Hawaiian Pacific Islander, Hispanic, Middle Eastern, Unknown, and White. I used the three most prominent racial subgroups (*White*, *Black*, *Hispanic*). My reference variable is White, because it is important to determine the effects of arrest rates on traditionally targeted communities with respect to the least targeted racial group. I use dummy variables for both Hispanic and Black with respect to white. For *Gender*, the raw data contains male, female, and unknown. I filtered the data to include only male and female, and used female as my reference variable, due to males composing a larger percentage of the arrested population in Austin. Regarding *Race Known Before Stop*, Yes=1 and No=0.

I expect similar outcomes in both regressions. First, *Race* will significantly affect the dependent variable, but *Gender*, *Age*, and *Race Known Before Stop* will not affect the dependent variable. Second, the predictive model will be weak and ill-fitting because of the large N value my selectivity of independent variables. By adding more variables, I can increase the strength and fit of the predictive model, as indicated in Multiple R and Adjusted R Squared. I however will restrict the independent variables to those that indicate demographic factors that can impact arrests.

My null hypothesis is the same for both regressions. I anticipate that both *Location* and *Time* will be influenced by *Race*, *Age*, *Gender*, and *Race Known Before Stop*, with *Race* serving as the most influential factor. If either linear regression indicates that demographic factors influence the time or location of arrest, I can then reject the null hypothesis and assume correlation.

$$H_0: Age + Race + Gender + Race\ Known\ Before\ Stop = 0$$
$$H_1: Age + Race + Gender + Race\ Known\ Before\ Stop \neq 0$$

Results

Excel's multiple linear regression output gives information determining the value of the model and the effect of independent variables on the dependent variable, which is *Location* for the first regression and *Time* for the second. The coefficient of correlation (Multiple R) and the coefficient of determination (Adjusted R Squared) are indicators of predictive model fit and strength respectively. A coefficient of correlation equal to 1 indicates that the data perfectly fits the predictive model. A coefficient of determination of 1 indicates that 100% of the change in the dependent variable is explained by the independent variables. Large data sets generally have low R values, especially when the regression analysis utilizes only a few of all available variables. Adding more variables to the regression will increase R values, but not actually indicate valuable information regarding interpretation of the data.

Regarding each independent variable, I will consider three indicators of correlation. The first is the probability value. A p-value less than or equal to 0.05 signifies that the values observed in the regression will be observed in real life 95% of the time. Therefore, all independent variables with p-values less than or equal to 0.05 are considered significant in that there is an acceptably high probability that it affects the dependent variable. Second, the value 0 cannot fall within the 95% confidence interval. Third, the t-statistic for each significant variable must be less than -2 or more than 2. T-stat indicates the extremity of the observation. Independent variables with t-statistic values within -2 to 2 for sample sizes greater than 30 are unlikely to affect the dependent variable. Thus, all significance variables must have a t-statistic that falls outside the designated range.

In the first regression, where *Location* is the dependent variable, *Age* and *Race Known Before Stop* are significant, while *Gender* and *Race* are not significant. Significant independent variables only slightly affect the dependent variable, given that their coefficients are extremely small. The p-value, coefficient, and t-statistic for *age* are $1.68x10^{-2}$, $-9.44x10^{-5}$, and -2.39 respectively. *Age* is significant because the p-value is less than 0.05 and the t-stat is less than 2. Its coefficient indicates a negative relationship with *location*, the dependent variable. The p-value, coefficient, and t-stat for *race known before stop* are $5.19x10^{-3}$, $3.89x10^{-3}$, and 2.80 respectively. *Race known before stop* is significant because the p-value is less than 0.05 and the

t-stat is more than 2. Its coefficient indicates a positive relationship with *location*, the dependent variable. Neither independent variable's 95% confidence interval contains the value 0. Multiple R, the coefficient of correlation, is $5.05x10^{-2}$, indicating that the predictive model fails to explain nearly 95% of the variation in the outcome. The predictive model also lacks strength as indicated by Adjusted R Squared, the coefficient of determination, which is $1.73x10^{-3}$. The F-Statistic is 3.09 with an F-Table value of $8.56x10^{-8}$, indicating that the regression excellently explains the variability between dependent and independent variables.

Regression #1	Multiple R	R Square	Adjusted R Square	Standard Error	Observations	F Statistic	F Table Value	
	0.05052122	0.002552394	0.001727784	0.031727945	6054	3.095275756	0.008565879	
Variables	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Location	0.121120059	0.001549706	78.15679535	0	0.118082083	0.124158035	0.118082083	0.124158035
Age	-9.44469E-05	3.94986E-05	-2.391146878	0.016826217	-0.000171878	-1.70156E-05	-0.000171878	-1.70156E-05
Hispanic	0.000466118	0.000950151	0.490572228	0.62374683	-0.001396517	0.002328752	-0.001396517	0.002328752
Black	-0.00096986	0.001070226	-0.906219799	0.364855627	-0.003067883	0.001128164	-0.003067883	0.001128164
Male	-0.000691166	0.000960265	-0.719765854	0.471696984	-0.002573627	0.001191295	-0.002573627	0.001191295
Race Known before Stop	0.003891826	0.001392001	2.795851236	0.005192654	0.001163009	0.006620644	0.001163009	0.006620644

$$Predictive\ Model\colon\ Location = 0.1\ - 9.44x10^{-5} * Age + .004 * RaceKnownBeforeStop$$

In the second regression, where *time* is the dependent variable, *gender* and *race* are significant, while *age* and *race known before stop* are not significant. Significant independent variables largely affect the dependent variable, given that their coefficients are relatively large. *Race* is *Hispanic* and *Black* with *White* acting as the reference variable. The p-value, coefficient, and t-statistic for *Hispanic* are $2.98x10^{-2}$, 22.45, and 2.17 respectively. *Hispanic* is significant because the p-value is less than 0.05 and the t-stat is more than 2. Its coefficient indicates a positive relationship with *time*, the dependent variable. The p-value, coefficient, and t-statistic for *Black* are $5.04x10^{-17}$, 97.85, and 8.41 respectively. *Black* is significant because the p-value is less than 0.05 and the t-stat is more than 2. Its coefficient indicates a positive relationship with *time*, the dependent variable. The p-value, coefficient, and t-statistic for *Male* are $1.23x10^{-2}$, 26.14, and 2.50 respectively. *Male* is significant because the p-value is less than 0.05 and the t-stat is more than 2. Its coefficient indicates a positive relationship with *time*, the dependent variable. None of the independent variables' 95% confidence interval contains the value 0. Multiple R, the coefficient of correlation, is 0.123, indicating that the predictive model fails to explain nearly 88% of the variation in the outcome. The predictive model also lacks strength as indicated by Adjusted R Squared, the coefficient of determination, which is $1.40x10^{-2}$. The F-

Statistic is 18.16 with an F-Table value of $6.33x10^{-18}$, indicating that the regression excellently explains the variability between dependent and independent variables.

Regression #2	Multiple R	R Square	Adjusted R Square	Standard Error	Observations	F Statistic	F Table Value	
	0.12160283	0.014787248	0.013972754	344.9158958	6054	18.15511979	6.32991E-18	
Variables	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Time	602.568703	16.84692315	35.76728507	2.4743E-254	569.5427311	635.594675	569.5427311	635.594675
Age	0.727382005	0.429390914	1.693985552	0.090319518	-0.11437718	1.56914119	-0.11437718	1.56914119
Hispanic	22.44516313	10.32913466	2.172995499	0.029819345	2.196378896	42.69394737	2.196378896	42.69394737
Black	97.85207073	11.63447087	8.410530383	5.04049E-17	75.04436244	120.659779	75.04436244	120.659779
Male	26.14433664	10.43908227	2.504466961	0.012289716	5.680015932	46.60865736	5.680015932	46.60865736
Race Known before Stop	-27.97045813	15.13249956	-1.848369994	0.064597619	-57.63554902	1.694632764	-57.63554902	1.694632764

$$Predictive\ Model:\ Time = 602.6 + 22.4 * Hispanic + 97.9 * Black + 26.1 * Male$$

I reject the null hypothesis in both instances. Location and Time are both influenced by composite factors embedded in the predictive models. By using a p-value of 0.05 for both regressions, I minimize the probability of error. Because I reject the null hypothesis, there is a possibility for Type 1 error if in fact the null hypothesis is true. However, in 95% of situations both predictive models will hold true, indicating significant correlation between dependent and independent variables.