STAT 4630: Statistical Machine Learning
Semester Project Final Report

Group 6
Eva Bogdewic, ejb4nd
Tim Kim, tsk8va
Corey Runkel, cnr3cg

Introduction

An article by ProPublica entitled "How Chicago Ticket Debt Sends Black Motorists into Bankruptcy" describes a phenomenon in which people in Chicago are exploited in order to boost the city's revenue. According to Sanchez and Kambhampati, traffic citations too often result in thousands of dollars worth of debt and even bankruptcy.

Question 1: Can the amount a motorist pays towards a parking ticket be predicted based on factors such as traffic stop location and driver origin?

The amount that a motorist pays cannot be predicted based on where the ticket was issued, casting doubt on Chicago's ability to identify and extract revenue from its lower-income motorists. Additionally, we were not able to accurately predict the final fine amount. However, several significant factors were identified. The best predictor of the total payment was, understandably, the initial fine amount. Following the initial fine amount, the ticketing authority turned out to be a significant predictor of the final fine amount. The data available could not encapsulate all potential confounding variables, but it allowed us to identify a few of the factors affecting the outcome of traffic citations in Chicago.

Question 2: Can the ultimate legal outcome of a ticket be predicted by factors such as amount of late fees, citation location, or reason for issuing the ticket?

For this second question, our results are not so much predictive as inferential. Firstly, the City of Chicago lacks the ability to extract revenue at will from its drivers. The City cannot determine from the location, time, make, or even the imposed fine whether or not the offender will eventually pay. However, insights ascertained from the data indicate that certain areas of Chicago, as evidenced by the high significance measurements in the logistic regression models, are paying more into the pocket of the city government, certain agencies are more effective at collecting, and out-of-staters who live far from Illinois may be less inclined to return to the state to contest their citations.

The initial claims made by ProPublica point toward a flawed system that fails to consider the impact that its provisional policies might have on some of its least represented groups. This project is interesting and important as it examines (a) groups of people that are disproportionately affected by these policies, whether the city intends this or not, and (b) how legal proceedings are impacted by these distinguishing factors.

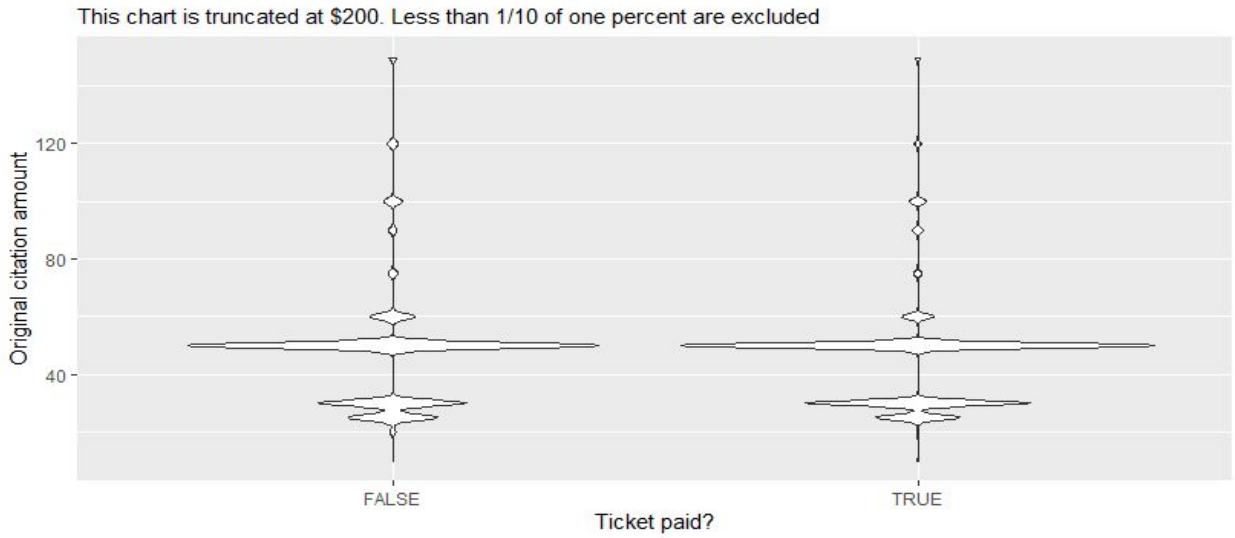
Data Processing & Cleaning

a. A description of any data cleaning your group had to do to prepare this data set. This could include removing certain observations, collapsing classes, categorizing quantitative variables, etc. Reasons should be provided why these data cleaning steps were taken. If your group did not have data cleaning to do, you can skip this section.

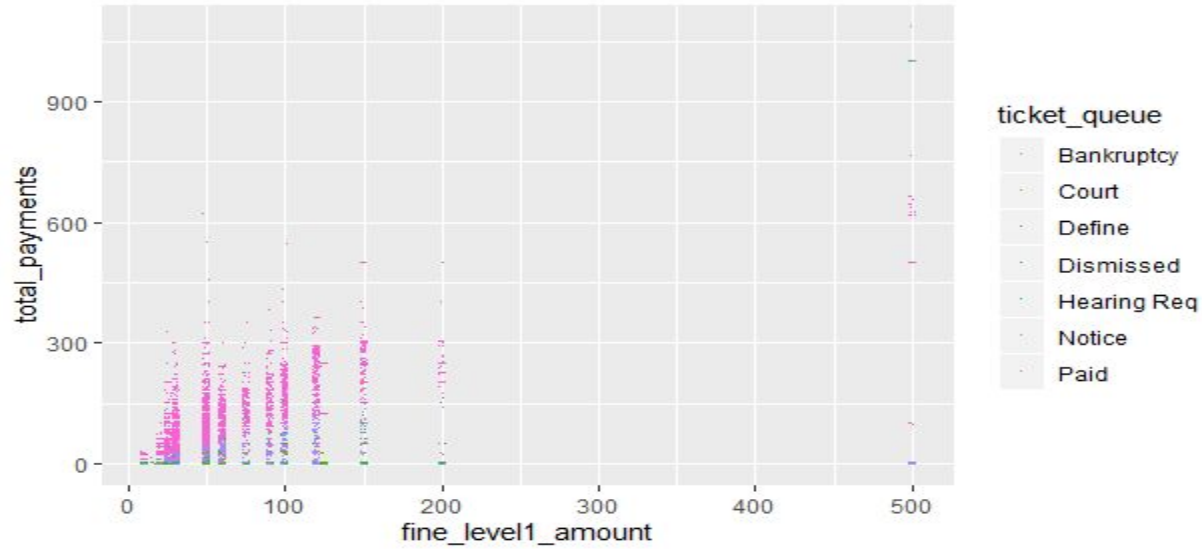
Our data cleaning fell primarily into two categories: cleaning for computation and cleaning for interpretation. For computation, we cut the original dataset of 28 million rows down to just under one million rows, as R originally demanded over 20GB of RAM. Additionally, a number of our models could not make use of `NA` values. We therefore used `na.omit()` to remove those observations.

For interpretation, our cleaning was extensive. We collapsed `community_area_name`, `vehicle_make`, and `state` into "meaningful" categories. The 76 official community areas in Chicago were translated into its 9 *de facto* official sides, such as the famous North and Southsides. This made our output comprehensible. To the best of our ability, we categorized vehicle makes according to how a ticketing authority might expect the willingness to pay of that car's driver. For instance, a Ferrari would imply a high income such that even $500 fines would be worthwhile to pay. There isn't quite an antonym for high-income cars, as a wide range of people buy Hondas, Fords, and the like. But it is enough that luxury cars and commercial vehicles signal a willingness to pay. Additionally, we grouped `state`--the state listed on a license plate--into three classes: `Illinois`, `bordering`, and `Other`. The logic behind this collapse is that drivers will not make an additional trip just to contest a parking ticket, where the cost of travel may already outstrip the cost of the ticket itself. Lastly, we transformed `hour` and `month` because we suspected seasonal effects. Chicago's tourism waxes in the summer, leading to more drivers with less experience parking in the City, and wanes in the winter. During the day, we suspected that late-night or midday parking tickets may differ from tickets issued just before and after workdays. For these reasons, we added 2nd-order transformations of these predictors.

Exploratory Data Analysis



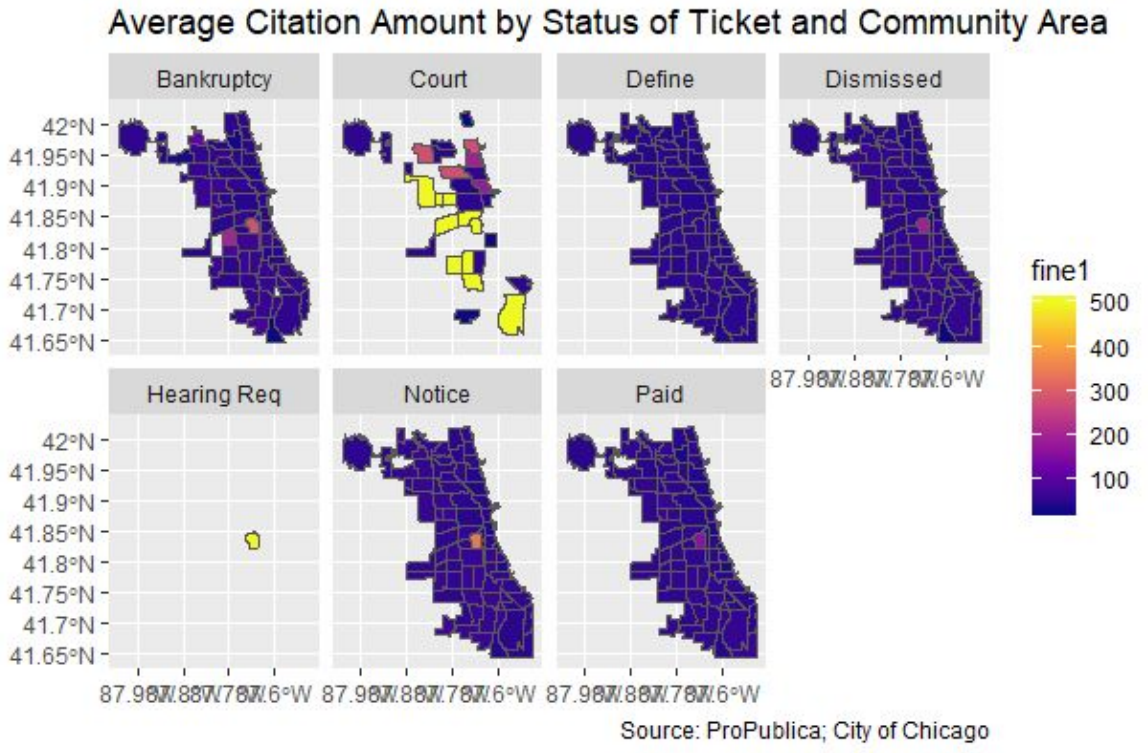
This violin plot highlights two characteristics of our data. First, it reinforces that tickets are issued in discrete amounts, not on a continuum. This discreteness suggests that linear regression may not perform as well as methods offer discrete response variables. Second, the differences between the original citation amount in paid and unpaid tickets appear to be negligible. This suggests that it will be difficult to predict `extract`, though it is a discrete variable, using `fine_level1_amount`, our most important predictor for question 1. Furthermore, the negligible differences do not rule out the claim that similar laws were cited for paid and unpaid tickets, since the amounts were the same. Note that there are several citations that full under a single citation amount.



While tickets are not issued on a continuum, they end up with far more values as total payments. This fact may suggest that we ought to have mined the fees and additions to tickets for discriminatory behavior. More directly important to answering out question,

however, is the approximately 45-degree line over which tickets are almost-universally `Paid`. This scatterplot illustrates the importance of `fine_level1_amount` in determining `total_payments`.

Below maps the 76 community_area_names and their average citation amount in each of the 7 legal categories a ticket could be in. Three findings are of note. First, the Bridgeport community--shaped vaguely like Wisconsin and the only neighborhood visible in the Hearing Req map--is the only community with any persistent differences. In `Hearing Req`, its uniqueness is due to the uniqueness of the underlying data: there is a single observation in `Hearing Req`. The reasons for persistent differences across legal statuses are completely unknown; searching online for Bridgeport and its parking yields no leads. Second, notice how heterogenous the results for Court are. While there were over 650,000 Paid tickets, only 78 were tied to a court case, letting the [ir]rationality of legal opinion disperse the payments across fine levels. Conversely, notice how pedestrian most of these community areas are. In all but court, nearly every community area exhibits the same behavior.

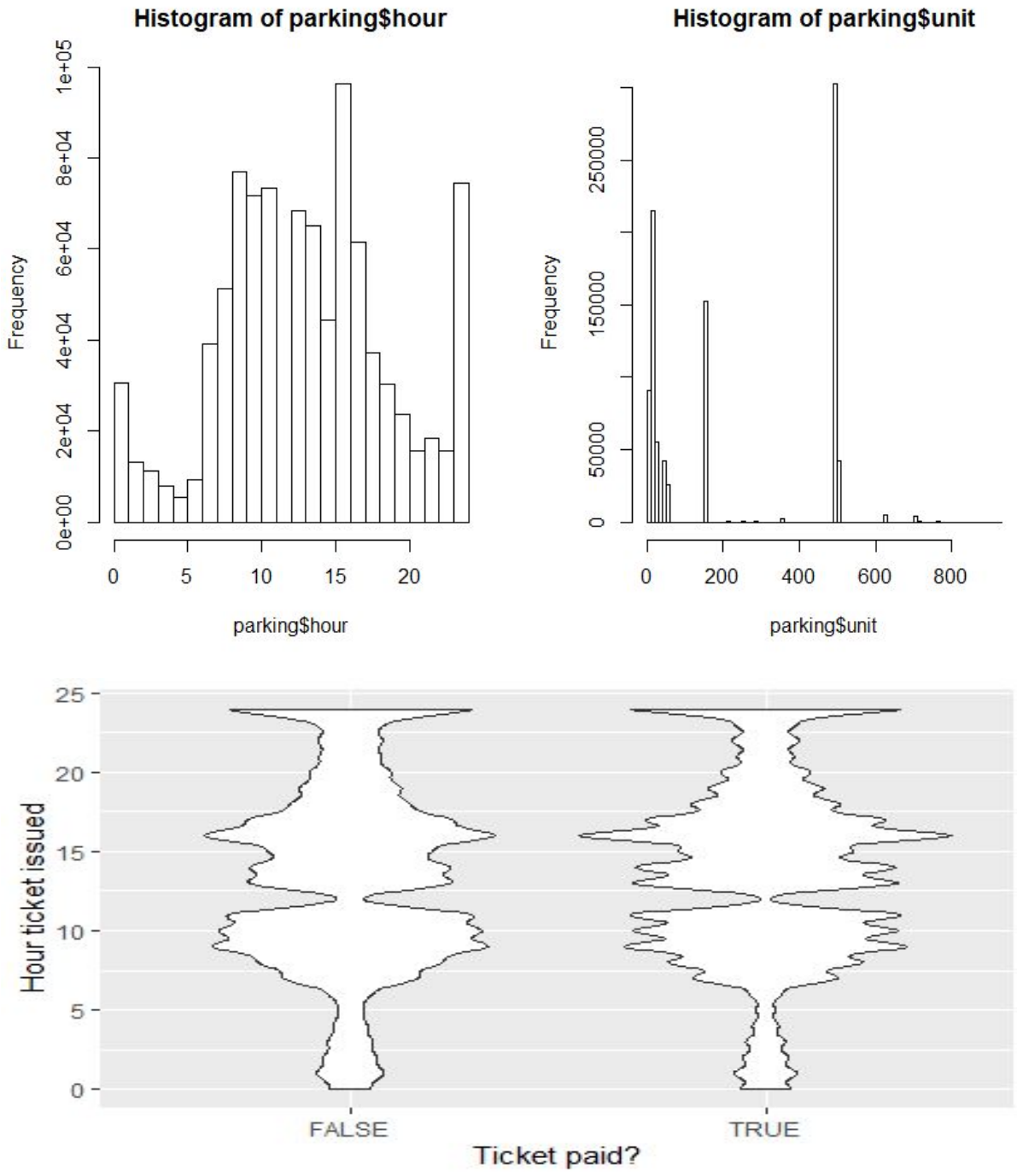


Examining the different wards of Chicago in the context of this problem motivated us to split the city into nine "sides." The 76 different community areas were too many, both from a computational and interpretational perspective. For this reason, we consolidated the community areas into the `side` variable, corresponding to Chicago's official "sides": `Central`, `South`, `Southwest`, `Far Southeast`, `Far Southwest`, `North`, `Northwest`, `Far North`, and `West`. This consolidation reflects historical and political realities, and only one addition and one split has occurred in about 90 years. As **R**

defaults to alphabetizing factors, `Far North` is the reference class. It features part of Northwestern's campus and is affluent compared with most of the city, though not quite so much as the Central side, with the Magnificent Mile and the Gold Coast.

Many of the variables contained too many categories to make the results of the regression significant or conducive for analysis and inference. For example, while we initially set out to compare the different units within the Chicago Police Department and private contractors, we realized that there were hundreds and that it would be more meaningful to compare those citations falling under the jurisdiction of the Chicago Police Department, those falling under the jurisdiction of the Department of Finance, and those categorized as "miscellaneous."

Comparison of the frequency of citations recorded by the hour drew our interest to the hour variable although the final cost did not seem to vary much across the hours. Similarly, we decided that instead of comparing the citations across states, we would compare in-state citations against bordering state citations.



Analysis from Regression, Classification, Trees

Linear Regression

The regression model we created seems to adequately predict the amount a fined motorist would pay, with an R^2 value of 0.6183. Furthermore, R's significance

measure of the variables we chose to include indicates very strong significance from 15 variables. The diagnostics of the model tell a more nuanced tale: within the bounds where most observations fall, our model does an excellent job. However, there are well-separated outliers that require a different treatment than simply linear regression to account for.

Something of interest was the influence that "miscellaneous" police units had on tickets. The unit category "Miscellaneous" was assumed to represent private security companies, and it was found that the tickets they issued were comparatively small. Most people do not readily associate private contractors with parking tickets, and it could be interesting to further investigate how much impact these miscellaneous units have on issuing parking tickets. Another note of interest was the general uniformity of ticket costs. Despite our dataset including nearly a million data points, the ticket costs and payments only really had around 100 distinct points. This kind of extremely discretized data could be worth further investigation, and our classification exercises will help to do just that.

Ultimately, the results of our model-building gave us some of the answers that we were searching for in our thought experiment. For instance, `season2` has a significant positive value. This means it opens up, and thus that the winter sees higher ticket prices than the summer, after controlling for other factors. Also, note that tickets that go to court have much lower parking tickets. This contrasted somewhat with the results of our first regression, which included the [dropped] variable state. That regression associated out-of-state plates with similar ticket prices to in-state license plates. Of course, the data available could not encapsulate all potential confounding variables, but it allowed us to identify a few of the factors affecting the outcome of traffic citations in Chicago.

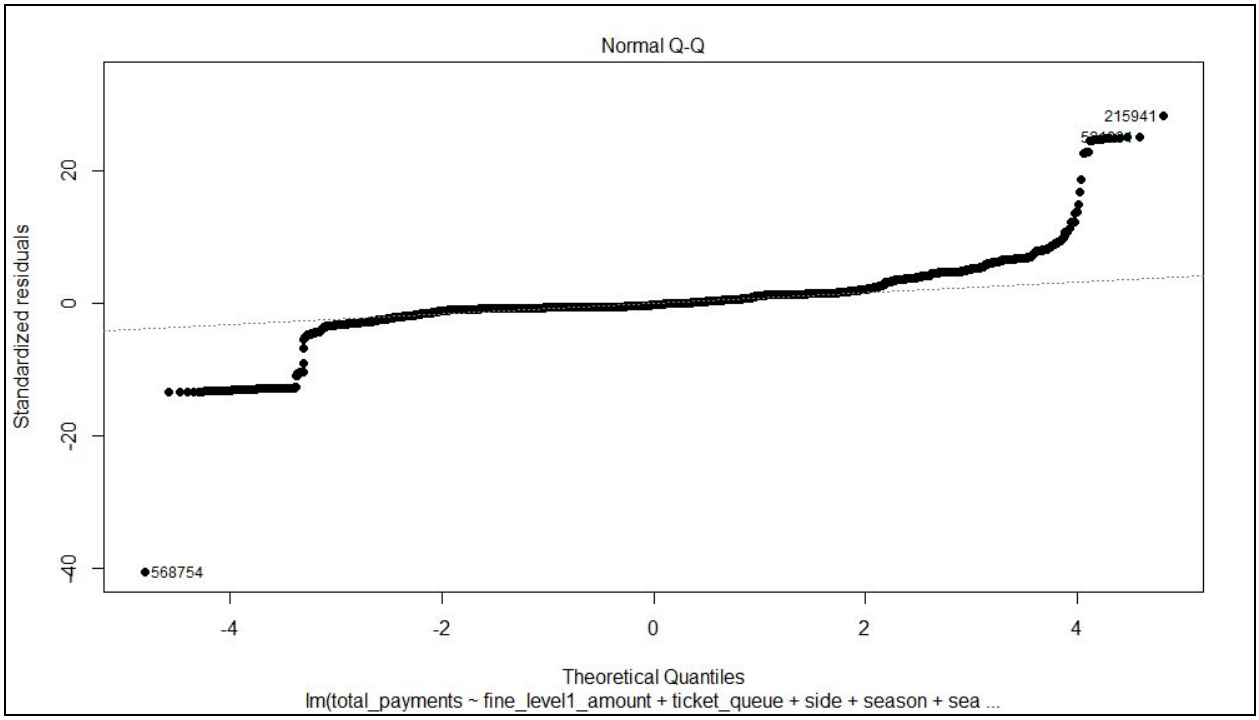
Overall, our linear regression performed well in the middle of the dataset,

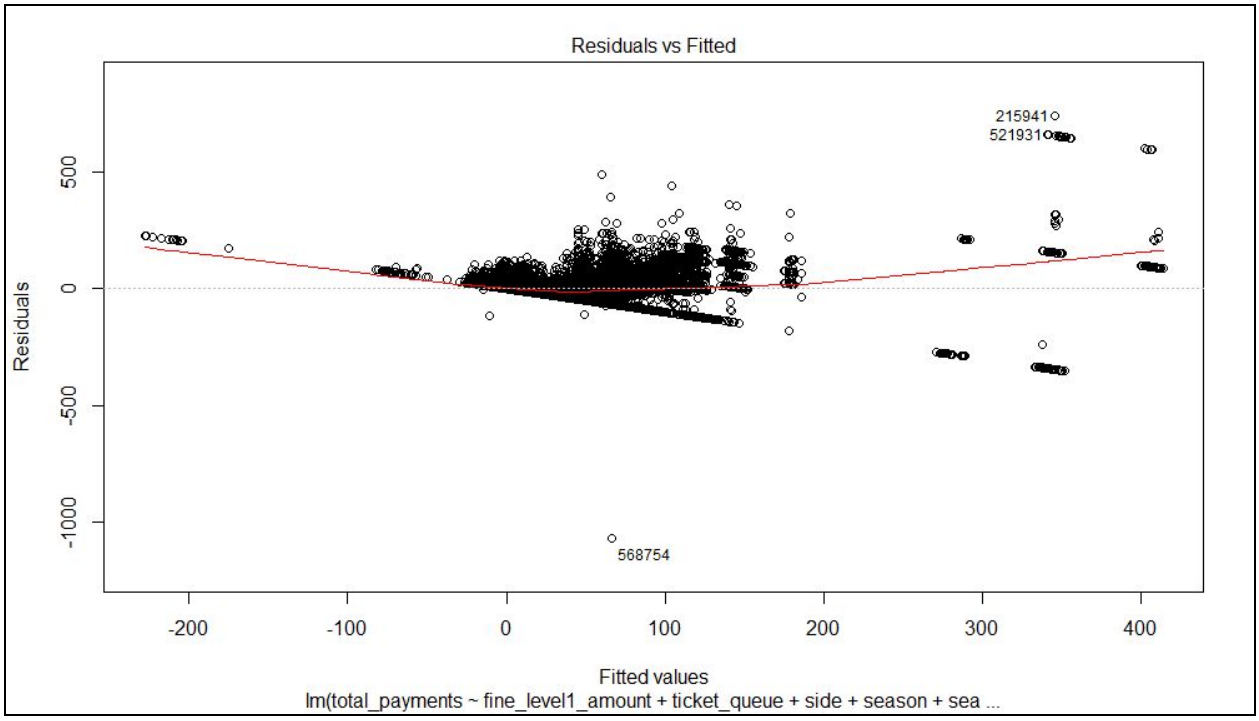
Summary() call on linear regression model with factors:

```
Call:
lm(formula = total_payments ~ fine_level1_amount + ticket_queue + side + season +
season2 + unit_description, data = parking.train)

Residuals:
     Min       1Q       Median       3Q      Max
   -1066.28   -16.06    -6.46    12.51   739.63

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -4.085e+01  5.112e-01 -79.911  < 2e-16 ***
fine_level1_amount            7.492e-01  1.440e-03 520.256  < 2e-16 ***
ticket_queueCourt            -2.060e+02  3.605e+00 -57.147  < 2e-16 ***
ticket_queueDefine            3.153e+00  5.021e-01   6.280  3.4e-10 ***
ticket_queueDismissed         1.241e+01  5.086e-01  24.400  < 2e-16 ***
```

```
ticket_queueHearing Req       -3.369e+02  2.634e+01 -12.793  < 2e-16 ***
ticket_queueNotice            -1.481e-01  5.061e-01  -0.293   0.7698
ticket_queuePaid               6.693e+01  4.972e-01 134.633  < 2e-16 ***
sidesouthwest                  2.831e+00  1.781e-01  15.894  < 2e-16 ***
sidesouth                      3.607e+00  1.753e-01  20.575  < 2e-16 ***
sidefarsouthwest               1.055e+01  3.831e-01  27.546  < 2e-16 ***
sidewest                       4.993e+00  1.216e-01  41.061  < 2e-16 ***
sidefarsoutheast               8.703e+00  3.488e-01  24.954  < 2e-16 ***
sidenorth                      4.188e-02  1.308e-01   0.320   0.7489
sidenorthwest                  4.944e+00  2.614e-01  18.911  < 2e-16 ***
sidecentral                    3.199e+00  1.126e-01  28.418  < 2e-16 ***
season                         5.462e-01  1.137e-02  48.040  < 2e-16 ***
season2                        8.207e-02  3.605e-03  22.765  < 2e-16 ***
unit_descriptionDOF           -2.655e+00  8.038e-02 -33.033  < 2e-16 ***
unit_descriptionMiscellaneous -4.741e-01  1.664e-01  -2.848   0.0044 **
unit_descriptionUnidentified  -6.243e+01  1.877e+00 -33.260  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.32 on 658898 degrees of freedom
Multiple R-squared:  0.6183,    Adjusted R-squared:  0.6183
F-statistic: 5.337e+04 on 20 and 658898 DF,  p-value: < 2.2e-16
```

Below are two diagnostic plots for the linear regression: a Q-Q plot and a residual plot:





Both the quantile-quantile and residuals plot show that the model fits a majority of the data, but that it performs poorly at the edges. This dataset's discreteness makes the very edges of the model perform particularly poorly. These are not ideal assumptions for OLS, and the model's results make clear that fact.

Logistic Regression

Adding the categorical predictors “state”, “side”, and “unit description” to the logistic regression model slightly improved its predictive power. The calculated AUC was 0.6055759, slightly more accurate than that of the previous model, 0.5700081.

These categorical variables were all significant when exploring the previous question. We determined that, if significant in answering this question, these predictors would provide valuable insight about the procurement of traffic citations.

In the ultimate model with categorical variables, `fine_level1_amount`, `hour`, `month`, `state`, `side`, and `unit_description` are significant. Within the state variable, only the `stateOther` variable was significant. The coefficient of 3.54 means that, all other variables held equal, a car with an out-of-state, non-bordering license plate is 35 times more likely to pay a parking ticket than cars with Illinois license plates! Intuitively, this means that distant drivers are much less likely to spend the gas and time to contest a ticket.

Within the `side` variable, all categories are significant aside from `sidecentral`. The category `sidenorth` had a positive coefficient; all other significant categories had negative coefficients. This means that, holding all other variables equal, an observation labeled “sidenorth” is more likely to result in a “Paid” outcome while observations in the other areas will be less likely to result in a “Paid” outcome. It is not surprising that the sidecentral variable is less significant because it contains many observations so a relationship may be more difficult to detect.

Within the unit_description variable, the categories `unit_descriptionDOF` and `unit_descriptionMiscellaneous` are both significant. Both have positive coefficients, meaning that, holding all other variables equal, observations from these categories are more likely to result in “Paid” outcomes. It is not surprising that the “unit_descriptionUnidentified” is insignificant, as the unidentified category likely spans many different, unrelated observations.

The summary of the logistic regression model is shown below:

Summary() call on logistic regression model with factors:

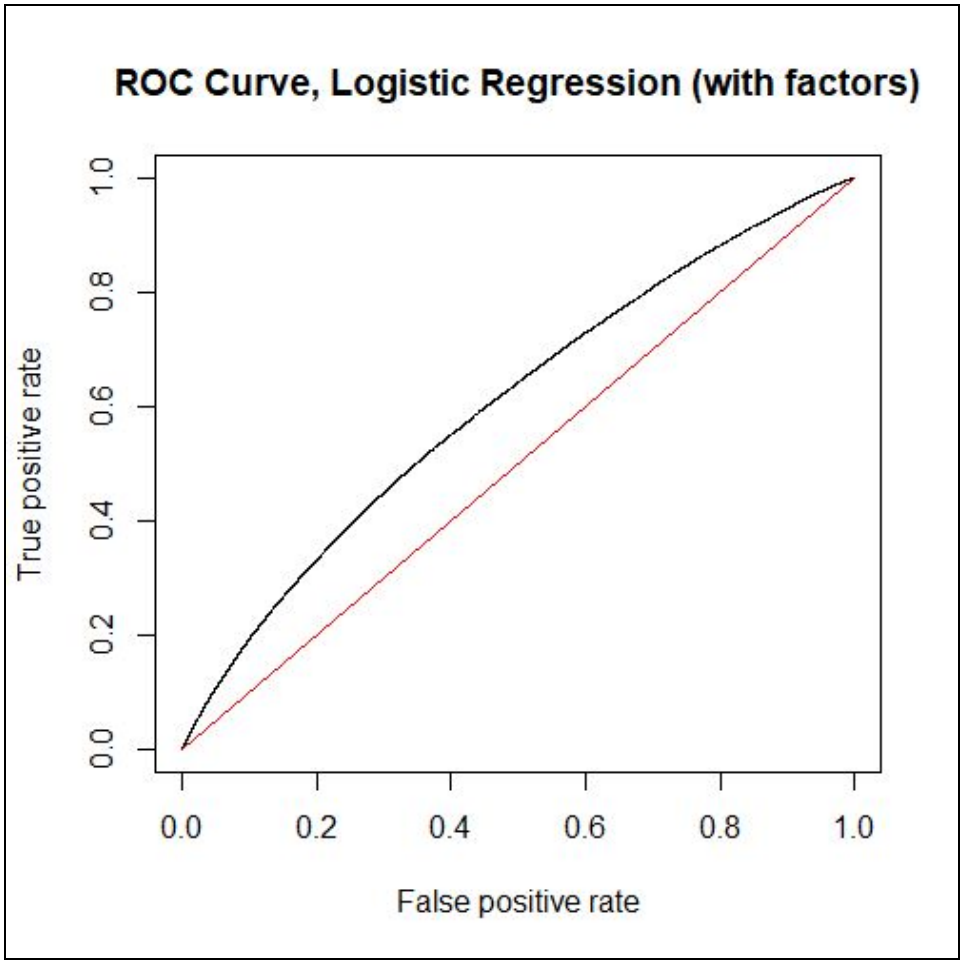
```
Call:
glm(formula = extract ~ fine_level1_amount + hour + hour2 + month + state + side +
unit_description, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8965  -1.3613   0.7874   0.9191   2.2223

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -2.499e+00  1.031e+00  -2.423 0.015392 *
fine_level1_amount          -3.626e-03  1.409e-04 -25.744  < 2e-16 ***
hour                         1.135e-02  2.060e-03   5.512 3.55e-08 ***
hour2                       -2.151e-04  7.497e-05  -2.869 0.004112 **
month                       -5.636e-02  1.110e-03 -50.788  < 2e-16 ***
statebordering               1.365e+00  1.034e+00   1.320 0.186795
```

```
stateOther                      3.540e+00  1.031e+00   3.432 0.000599 ***
sidesouthwest                  -3.958e-01  1.660e-02 -23.845  < 2e-16 ***
sidesouth                      -8.766e-02  1.702e-02  -5.151 2.60e-07 ***
sidefarsouthwest               -3.544e-01  3.574e-02  -9.915  < 2e-16 ***
sidewest                       -2.389e-01  1.165e-02 -20.514  < 2e-16 ***
sidefarsoutheast               -4.524e-01  3.239e-02 -13.969  < 2e-16 ***
sidenorth                       2.468e-01  1.310e-02  18.837  < 2e-16 ***
sidenorthwest                  -3.639e-01  2.437e-02 -14.928  < 2e-16 ***
sidecentral                     4.019e-04  1.090e-02   0.037 0.970577
unit_descriptionDOF             3.286e-01  8.111e-03  40.520  < 2e-16 ***
unit_descriptionMiscellaneous   9.163e-02  1.589e-02   5.766 8.10e-09 ***
unit_descriptionUnidentified    4.246e-01  1.869e-01   2.271 0.023146 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 589264  on 465257  degrees of freedom
R
esidual deviance: 576157  on 465240  degrees of freedom
  (5398 observations deleted due to missingness)
AIC: 576193

Number of Fisher Scoring iterations: 4
```

Below is the confusion matrix corresponding to the logistic regression model.

```
> table(test$extract, logit.preds.f > .5)

      FALSE  TRUE
FALSE 3667  148812
TRUE  2825  309900
```

The recall rate is 0.5648. The precision is 0.02404.



Let us be clear: our models lack predictive power. In classifying tickets into paid and unpaid categories, they fare little better than a coin flip. But we can draw two conclusions from the information offered by the models and their comparisons.

In regards to this question, our group was interested in the "Bankruptcy" class of the ticket_queue variable. However, given that there were only approximately 4,000 so categorized observations among more than a million observations, all models produced were difficult to interpret. The categorical variables included in our revised model are essential to answering our question of interest. This is a limit of the linear discriminant analysis approach.

Firstly, the City of Chicago lacks the ability to extract revenue at will from its drivers. The City cannot determine from the location, time, make, or even the imposed fine whether or not the offender will eventually pay. However, insights ascertained from the data indicate that certain areas of Chicago, as evidenced by the high significance measurements in the logistic regression models, are paying more into the pocket of the city government, certain agencies are more effective at collecting, and out-of-staters who live far from Illinois may be less inclined to return to the state to contest their citations. This supports both our theory

and the claims made in the ProPublica article that the blanket policy of increasing ticketing fines more strongly impacted poorer regions of Chicago.

Regression Tree

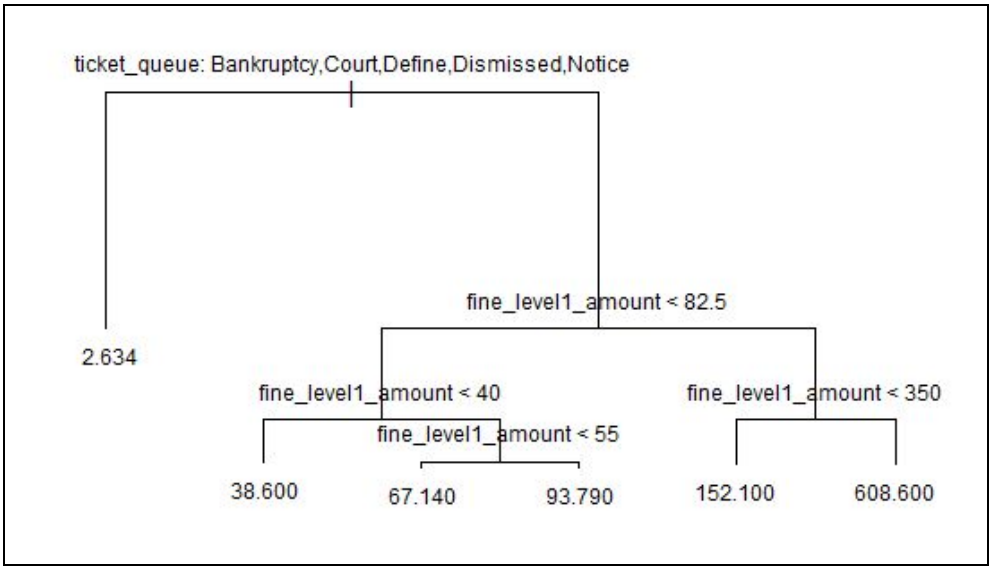
Some additional cleaning was performed before the regression tree was built. Missing value observations were omitted from the data. The `as.factor()` function was used to improve our interpretation of the `unit_description` and `ticket_queue` variables. The regression tree was built using the same regression model.

Regression tree:

```
tree(formula = total_payments ~ fine_level1_amount + ticket_queue + side +
season + season2 + unit_description, data = train)
Variables actually used in tree construction:
[1] "ticket_queue"      "fine_level1_amount"
Number of terminal nodes:  6
Residual mean deviance:  490.5 = 4614000 / 9407
Distribution of residuals:
      Min.  1st Qu.   Median  Mean  3rd Qu.     Max.
-108.600  -17.140   -2.634    0.000 1.396  391.400
```

Fig.4: *Summary of pruned tree*

The regression tree has **6 terminal nodes** and only **predictors `ticket_queue` and `fine_level1_amount`** were used. The graphical output is shown below:



Classification Tree

Classification tree summary output:

```
Classification tree:
tree(formula = extract ~ ., data = train)
Variables actually used in tree construction:
character(0)
Number of terminal nodes:  1
Residual mean deviance:  1.266 = 589200 / 465200
Misclassification error rate: 0.3286 = 152877 / 465230
```

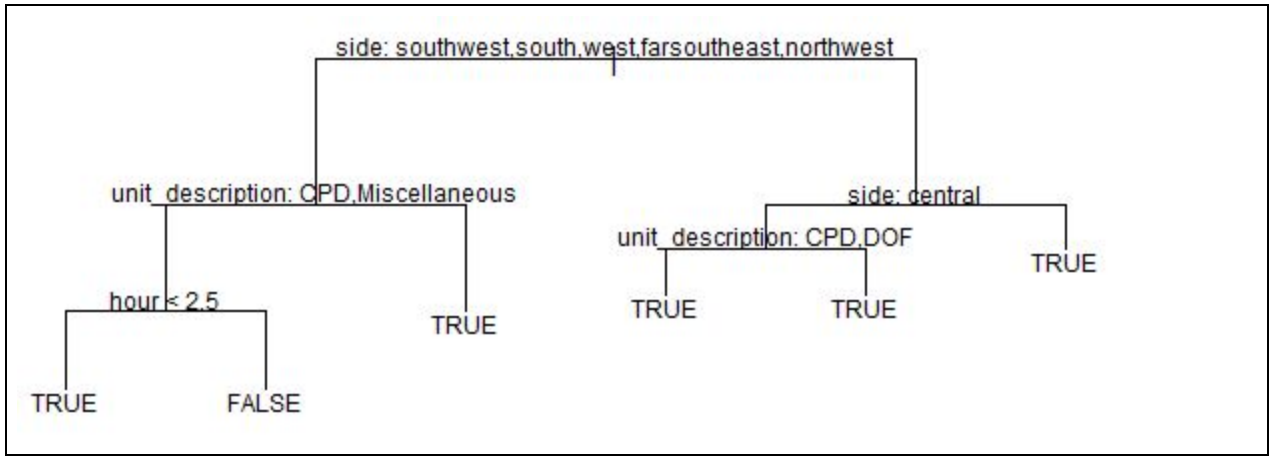
Fig. 5: *Erroneous but technically correct classification tree*

We can see that the technically correct tree has 1 terminal node and uses only the predictor `character`. Through extensive exploration, it was discovered that the tree could be improved by using a sampling size of 449 observations collected by random sampling.

The new classification tree has 6 terminal nodes and uses predictors `side`, `unit_description` and `hour`. Below are the R output and the classification tree pertaining to this improvement:

```
Classification tree:
tree(formula = extract ~ ., data = train)
Variables actually used in tree construction:
[1] "side"             "unit_description" "hour"
Number of terminal nodes:  6
Residual mean deviance:  1.167 = 524.2 / 449
Misclassification error rate: 0.2879 = 131 / 455
```

Fig. 6: *Summary of classification tree using smaller sample size*



We used bagging and Random Forests techniques to try to further improve the classification tree with the smaller data set. The R code is shown below and it indicates that neither of these techniques serves to improve the tree.

```
> importance(bagging)
                    %IncMSE   IncNodePurity
fine_level1_amount 117.357006    5707969.9
ticket_queue        303.252809   7546714.9
side                19.441770    490727.1
season              9.532682     388732.9
season2             6.540708     224367.1
unit_description    11.973619    429393.2
```

```
> importance(forest)
                    %IncMSE   IncNodePurity
fine_level1_amount  92.115837    4267837.5
ticket_queue        286.615570   7791528.7
side                11.631314    441919.7
season              10.076762    319107.5
season2             5.547383     199652.9
unit_description    9.871391     743136.7
```

Our regression tree was not of much use in predicting the response variable: the only variable that was used in the binary splitting was fine_level1_amount, and the response variable values at the nodes were discrete values that are not of much use in prediction.

As one can observe, the pruned tree is exactly the same as the one originally produced. The test MSE is `511.8`.

Bagging and randomForesting returned similar results, and posed similar challenges. Note in our code that we do not split the data into subsets of roughly equal sizes. Our dataset is far too large to run randomForest algorithms on personal computers. Even a 5% sample demanded a 7.0 GB array. With a 1% sample, our MSE were about 25 points higher than our pruned/original trees yielded. Still, they functioned.

Bagging returned an MSE of `508.8`, functionally similar than our pruned model. The random forest, however, returned an MSE of `498.4`, a clear improvement over our pruned model. Still, their importance calls told the same story as the pruned tree did: `ticket_queue` was by far the most significant factor in reducing deviance; then a massive plunge down to `fine_level1_amount`; then a slightly less-massive tumble to the rest of our predictors `side`, `season`, `unit_description`, and `season2` in that order. These results tell us the same story as the pruned tree: bureaucratic and legal factors decide far more of a ticket's final cost than do neighborhood, season, or ticketing authority.

In order to produce our classification tree, we had to take a very small sample of data from the original dataset. Pruning the tree yielded the same number of optimal terminal nodes; therefore no difference between unpruned and pruned trees.

Pruned error rate:	0.2879
Bagging error rate:	0.3618
Random forest error rate:	0.3597

Both bagging and random forest reported side and unit_description as the most important variables. Random forest placed considerable weight on side, with a mean decreasing accuracy of 11.13, compared to bagging's weight on the same variable of 7.89.

Any analysis of these results must be prefaced with the above challenge, in that all of the trees were generated using a 0.1% sample size; any samples with more observations yielded suspiciously sparse trees that classified everything as TRUE and did not utilize any variables at all. Hence, from this potentially unrepresentative sample, we cautiously conclude that side, which represents the community area that the vehicle was ticketed in, is the single most influential factor in determining whether the City of Chicago will manage to extract parking ticket fees from a motorist, followed by the unit the officer belongs to. This presents the somewhat troubling interpretation that certain areas of Chicago are more likely to issue parking tickets, especially given the context that the different community areas of Chicago are often marked by populations of varying socioeconomic composition. We can therefore answer our question of interest: We can predict the ultimate legal outcome of a ticket based on factors that an issuing officer would know, such as location and time of day.

Results

There were many similarities produced through our use of these different types of analyses. First and foremost, the models produced are not useful tools for predicting the financial or legal outcomes associated with Chicago parking citations. This result is expected for two reasons. First, if perfect practices are upheld by these parking authorities and the circumstances of those ticketed are similar, only the reason for ticketing, or type of ticket, should be a reliable predictor for the ultimate outcome of the ticket. Second, contrary to the first reason, ticketing has a level of random subjectivity because its process contains a human element.

The linear and logistic regression models both provided insight as to the significance of many factors associated with ticketing. Both types indicate that, as was reported in the article, poorer areas of Chicago are more heavily impacted by the citation practices than more affluent areas. In addition, apart from the obvious initial fine amount, the hour, month, and unit description were significant factors in predicting the outcomes of the citations. The regression and classification trees did not provide any additional insights.

Further Work

Much of our work was computationally limited; the original dataset was over 20 gigabytes in size, and it was infeasible for us to utilize all of this data. Being able to use all of this data could have offered us stronger statistical influences. Additionally, the response variable we originally wished to measure for our classification model was whether motorists were being pushed to bankruptcy by the policy changes. Unfortunately, by reducing our dataset, we did not have enough observations of “Bankruptcy” to produce substantial results.

In addition, we would be interested in a more detailed look at geographical variables. Combining our community areas into cardinal direction “sides” likely erased socio-economically distinct regions of Chicago that could have been of interest. Furthermore, the reference map we used to create the “sides” variable was demarcated in the 1920’s. We would be interested in using more recent data that would more accurately reflect how Chicago’s population has changed.

An interesting study might be to compare data from before and after the policy change, to look for a statistically significant increase in revenue from parking violations as well as increasing rates of bankruptcy induced by parking ticket debt.

Finally, a step we took in our data cleaning process was to combine the ticketing agencies into three groups. We are, however, interested in possible human biases, such as ticketing quotas by state police departments, or differences in professionalism and training between state deputies and private contractors. A more detailed look at the “unit” variable could yield interesting and noteworthy results.