

**Faculty of Information, University of Toronto**

**INF1344: Introduction to Statistics for Data Science**

**Professor Abbas Keramati**

Group 11:

Hamid Parsazadeh (997034472)

Isabel Bowman (1003469640)

Kuo-Lun Chang (1007618641)

Bethlehem Zebib (1008516660)

Ragave Vicknarajah (1003959610)

**Communicating Legal Precedent by Estimating Preparation Hours Needed for Criminal Cases**

**Estimating Preparation hours based on Expected Trial Days**

**December 12, 2021**

**TABLE OF CONTENTS**

1. **Introduction**

**1.1 Background**

**1.2 Research question and objectives**

1. **Methodology**

**2.1 Data Collection**

**2.2 Data Analysis Method**

1. **Statistical population Review**

**3.1 Initial data visualization**

**3.2 Outliers Identification**

**3.3 Population Distribution**

**3.4 Initial Hypotheses**

1. **Model Fitting**
   1. **Initial models comparison**

**4.2 Best model selection**

**4.3 Testing key assumptions**

1. **Conclusion**

**Abstract**

Managing serious criminal matters such as murder, sexual assault, or weapon offences are very difficult, and each case requires a different number of Trial Days to resolve. Lawyers need a certain amount of Preparation Hours to effectively represent the case in court, however as they are paid by the hour, they will rationally ask for the maximum amount possible in order to maximize their earnings. To mediate this and ensure there is no needless overspending, budget management committees require some rationale to authorize a reasonable amount of Preparation Hours for each case. They often turn to historical cases to grant Preparation Hours based on precedent, looking at the number of hours granted for past cases with a similar number of Trial Hours.

Our team aims to ease the work of such budget management teams by devising a statistical method to estimate the number of Preparation Hours that have been historically granted based on the number of Trial Hours for each case. This project would end with a program that could receive a new case’s ‘Trial Hours’, and output an estimated ‘Preparation Hours’, based on historical data. This would allow for the legal precedent to be communicated clearly and numerically to the committee, thus reducing bias.

**1. Introduction**

When a person is convicted of a serious crime such as murder, sexual assault, or weapon offences their case is most likely taken to court. When a matter is taken into court the criminal is usually in need of a lawyer who will need to learn the case “inside and out”. In order to thoroughly understand the case a lawyer needs a certain number of hours to prepare in order to effectively present the case in court. In order to prepare for court a lawyer may need to do the following things: initial investigation of every aspect of the case, preparing a plan to keep everything in order and easily accessible, gathering all evidence, requesting all important documents, preparing exhibits and demonstrative aids for use at trial, creating detailed outlines of direct testimony and cross-examination questions, creating opening remarks and forceful closing arguments, considering jury selection objectives, and preparing jury instructions (citation).

Lawyers are typically paid by the hour; therefore, they tend to ask for the maximum number of hours possible in order to maximize their personal earnings. However, it should be noted that there is no legally set amount of time for a lawyer to prepare for trial. The amount of preparation time can range anywhere from none to possibly several years (citation). This section directly relates to the time length of the trial. In order to mediate the number of preparation time that lawyers seek and to ensure there is no unnecessary spending, budget management committees require some rationale to authorize a reasonable amount of preparation hours for each case. They often turn to historical cases to grant preparation hours based on precedent, looking at the number of hours granted for past cases with a similar number of Trial Hours. It should be noted that section 11(b) of the Canadian Charter of Rights and Freedoms states that “any person charged with an offence has the right to be tried within a reasonable time” (citation). The purpose of section 11(b) is to make sure that the trials of those who are charged with crime are not prolonged or delayed, rather the number of days in trial should be as short as possible.

In order to keep track of the number of preparation hours that lawyers bill there are budget management teams at legal firms. Our team aims to ease the work of such budget management teams by devising a statistical method to estimate the number of preparation hours that have been historically granted based on the number of trial hours for each case. This project would end with a program that could receive a new case’s ‘Trial Hours’, and output an estimated ‘Preparation Hours’, based on historical data. This would allow for the legal precedent to be communicated clearly and numerically to the lawyer who will be representing said case, thus reducing bias.

In order to find a solution to the mentioned research question the following plan was devised and organised in this paper. Section one includes an introduction to the problem and the significance of the problem, research question and hypotheses. Section two presents the methods – framework, data collection, and modeling techniques that will be utilized in order to solve the research question. Section three contains the results of the analyses that were run. Finally, Section four concludes the findings of the research, highlights future directions, and states the limitations of the study.

**1.1 Background**

The purpose of this project was inspired by data from the Legal Aid of Ontario. Legal Aid Ontario is a publicly funded and publicly accountable non-profit corporation, responsible for administering the legal aid program (citation). Due to the availability of this data to the group and today’s political climate where many innocent individuals are being charged of crimes they did not commit; it is of importance to look into how many hours lawyers bill for a case in relation to the number of hours at court. Majority of lawyers decide to charge hourly rates as it is one of the most straightforward legal fee options. However, this can be a disadvantage to clients as lawyers sometimes do not adhere to the ethical obligations to his/her client in terms of the fees for their work (Moorhead, 2011). The paper by Moorhead (2011) focuses on the tension between the commercial imperative to make a profit and ethical obligation of lawyers to promote their clients’ interest over their own. Using data from a practitioner survey, claimant interviews, and existing data from the Department of Business Innovation and Skills, Moorhead (2011) was able to conduct both a quantitative and qualitative analysis. The quantitative data analysis consisted of computing the following descriptive statistics: minimum, maximum, mean, and median (Moorhead, 2011). The findings of the paper suggest that lawyers do in fact put their own financial interests before their clients, such that they tend to overcharge their clients (Moorhead, 2011). However, the paper also states that there are possible ways for lawyers to charge their clients on a more equitable and ethical basis.

These findings of lawyers being interested mainly on their financial gain has also been found in the study by Baker (2010). The paper utilized data from a survey of individuals who have experienced parental alienation (Baker, 2010). A particular question included in the survey that is of interest for the purposes of this paper was that the “attorney were adequately prepared for hearings”. Unsurprisingly, “only between 16 and 25 percent [of participants] reported that all attorneys were prepared for hearings...” (Baker, 2010). Additionally, it was found that the lawyers that the participants hired for the case became unavailable or difficulty to reach after they received the payment for their legal fees, or they ask for more money than originally agreed upon due to time needed for the case (Baker, 2010). This further confirms the findings from the paper by Moorhead (2011) that many lawyers are focused more on their financial interests over the interests of their clients. For the sake of the client, or more specifically the individual that was charged of a crime, it is important that they are being charged by their lawyer appropriately. It is not ethically correct for lawyers to be over charging their clients when majority of them are not adequately prepared for court even after giving their client the assumption that they are based on the legal charges.

**1.2 Research Question and Objectives**

This paper proposes multiple algorithms in order to get an optimal number of preparation hours based on an estimated number of trial days based on historical data collected from the Legal Aid of Ontario. The paper will utilize multiple data analysis methods taught in INF1340: Introduction to Statistics for Data Science to aid budget management teams in estimating the appropriate time a lawyer requires to prepare for a trial. The aim of the proposed project is to **find what the best algorithm is to estimate the number of required Preparation Hours based on the estimated number of Trial Days for a coming criminal case.**

The following hypotheses were derived from our research question: (1) there is a significant correlation between Trial Days and Authorized Preparation Hours, (2) there is a significant difference between the distribution of different criminal charge categories.

**2. Methodology**

**2.1 Data Collection**

The data collection process begun by all group members doing a search online of websites such as Kaggle, GitHub, and City of Toronto. The group members then brought together a dataset that they would like to analyze for the purposes of the final project. The final dataset that was agreed upon by all members was collected from Legal Aid Ontario by Hamid Parsazadeh. The particular dataset is focused on specific charges against the defendant (charge), the year (year), the gender of the defendant (gender), the number of days in court (court\_d), and the number of hours needed for court preparation (prep\_h).

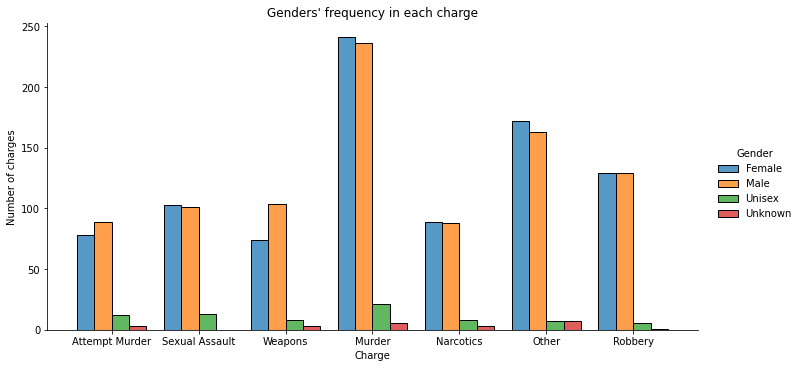
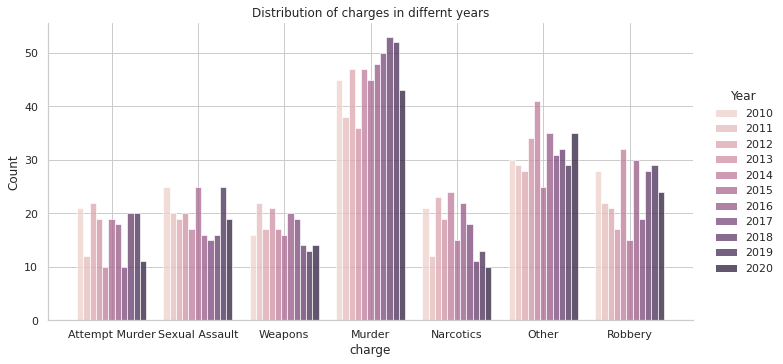
**2.2 Data Analysis Method**

Prior to creating our algorithms to solve our research question we summarized our data using multiple descriptive statistics methods. This includes creating multiple scatterplots to plot the preparation hours against the number of days in court based on the charge.

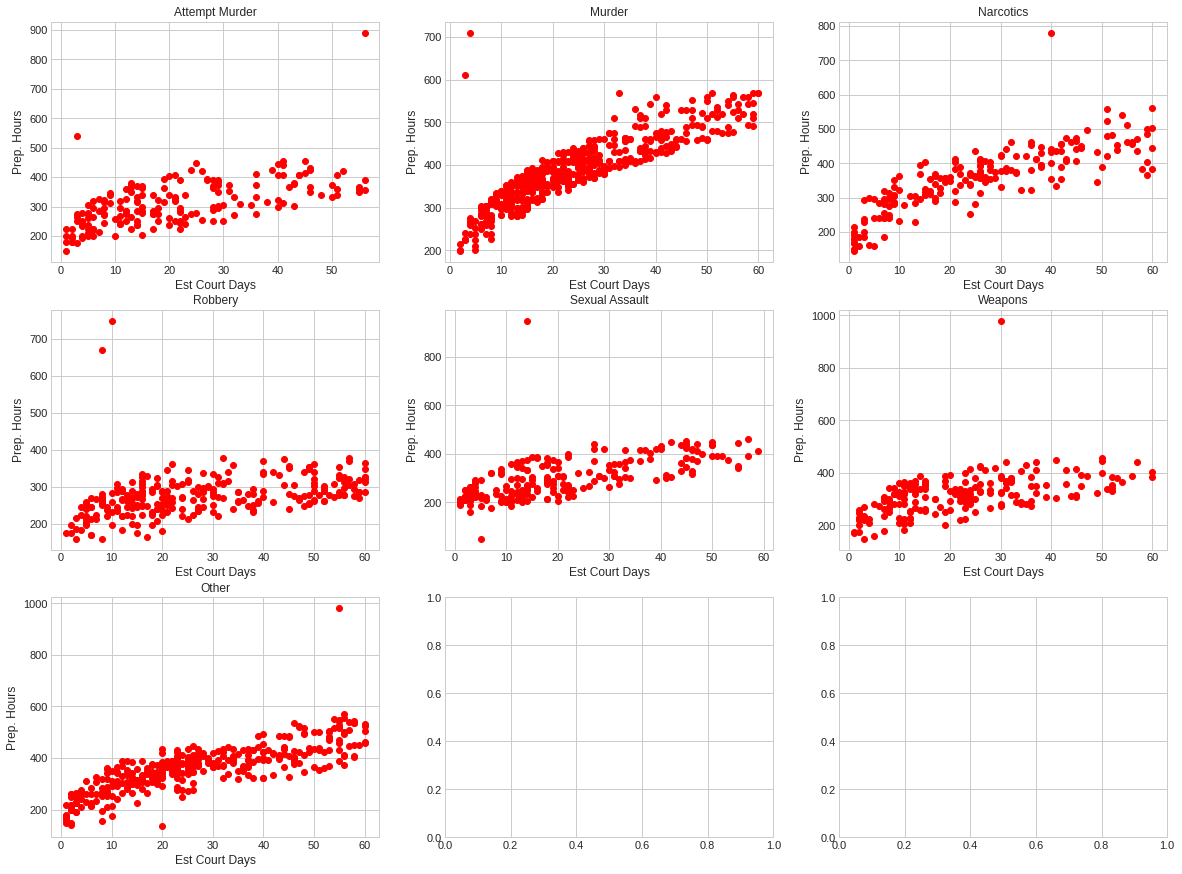
In order to test our research questions, we derived multiple algorithms in order to estimate the number of required Preparation Hours based on the estimated Trial Days for a coming criminal case.

**3. Statistical population review**

**3.1 Initial data visualization**

****

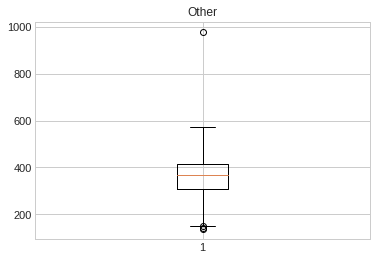
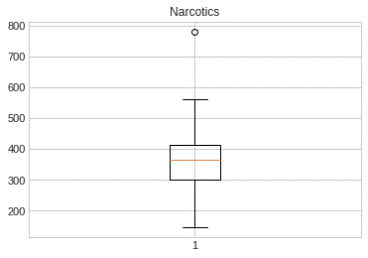
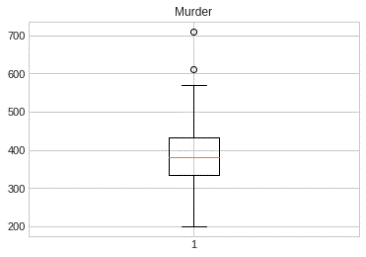
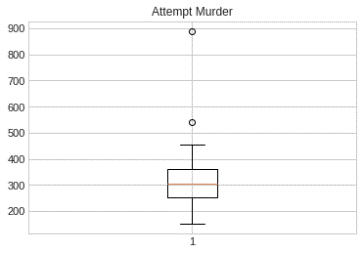
Is seems that the population is evenly distributed among different genders or over years.

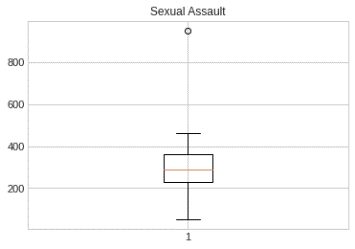
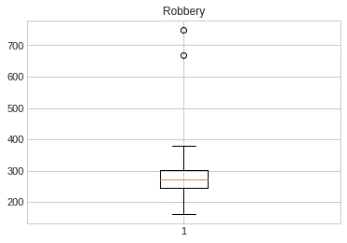
****

The scatter plots reveal that there is a trend between Est. Court Day and Total Preparation Hours. Moreover, some potential outliers are observed in different charges.

**3.2 Outliers identification**

We draw a boxplot for each charge and use the rule of “*1.5 times of the interquartile range*” to identify the outliers.





Here are the list of outliers:

**Gender Year Court\_d Prep\_h Charge**

Male 2010 3 540.0 Attempt Murder

Female 2013 56 890.0 Attempt Murder

Male 2020 3 610.0 Murder

Male 2017 4 710.0 Murder

Male 2015 40 750.0 Narcotics

Male 2020 1 148.0 Other

Female 2014 2 140.0 Other

Female 2018 2 150.0 Other

Female 2018 55 980.0 Other

Female 2011 20 135.0 Other

Male 2013 8 670.0 Robbery

Male 2018 10 750.0 Robbery

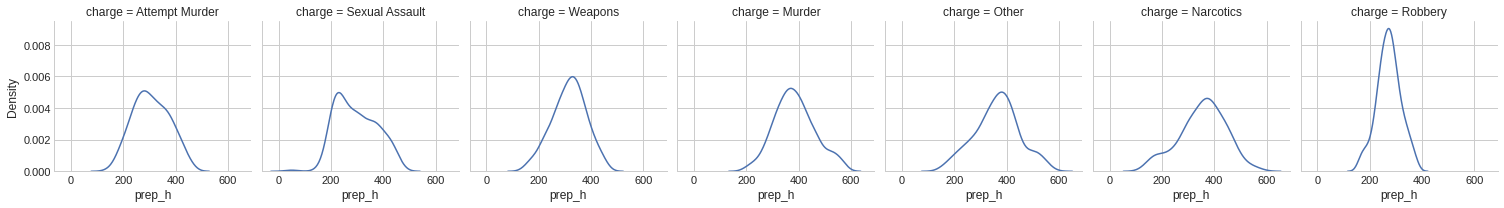
Female 2016 14 950.0 Sexual Assault

Male 2014 30 980.0 Weapons

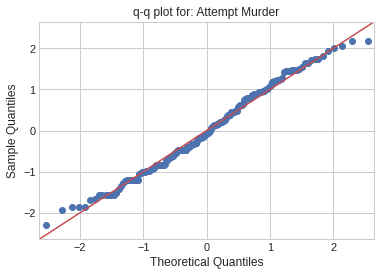
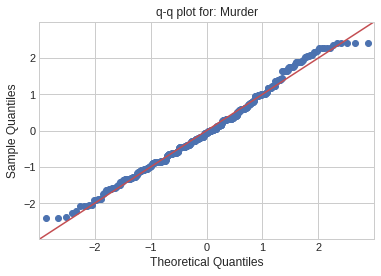
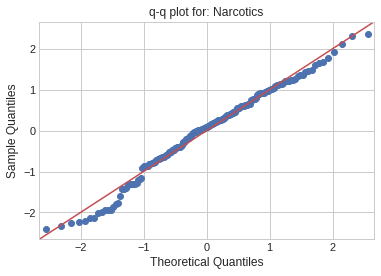
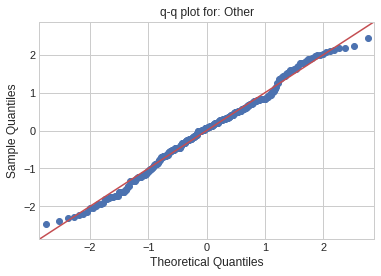
In discussion with the stakeholders, we can say that the outliers are due to very exceptional situations. Estimating preparation hour for these cases requires close coordination of Case managers with the acknowledged lawyers, so they always need to be managed manually and can be removed from the dataset to prevent biasing the final model.

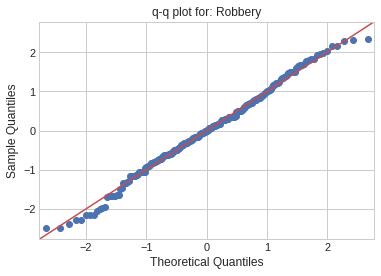
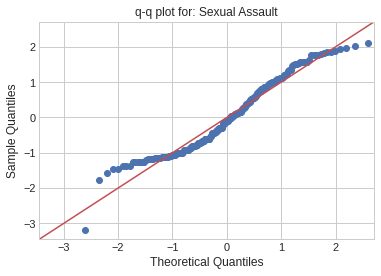
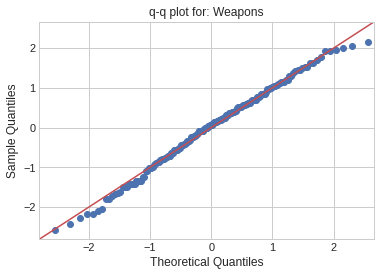
**3.3 Populations’ distribution**

Actual distribution of each charge:



Q-Q plot for each charge:

The Q-Q plots show that for all the charges, the points are forming a roughly straight line. It means that both Theoretical and sample Quantiles are came from the same distribution, so we can conclude that the population of the charges have Normal distribution. In order to verify this finding we use one sample kolmogorov-smirnov test, known as One Sample K-S test.

Using Python to perform test, here are the p\_values for each charge:

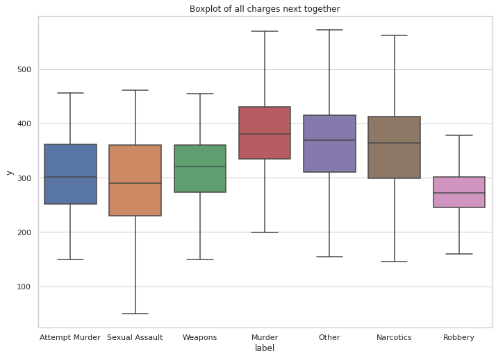
Attempt Murder: 0.48 Robbery: 0.75

Murder: 0.16 Sexual Assault: 0.02

Narcotics: 0.33 Other: 0.31

Weapons: 0.75

With α = 0.01 we can conclude that we do not have significant evidence to reject the Null Hypothesis for all of the charges. In continue we draw the Boxplot of each charge in one graph to have a visual comparison of the population distributions.



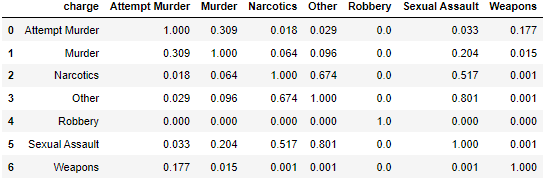
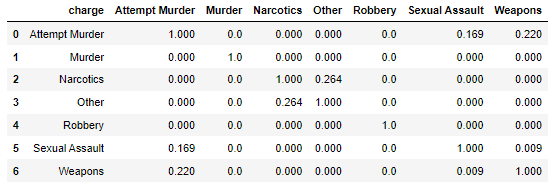
The Boxplot shows that some charges may have similar distributions. Since we have already verified the Normal distribution of all the charges, we will need to run the proper Two Independent Sample Tests for comparing the µ and σ of different charges (two by two) to discover the similar distributions or verifying that they are coming from different distributions. Therefore, we will have two sets of hypotheses:

and

Using Python, we run Leneve Test and T-Test for comparing variances and means of each set of two charges. [[1]](#footnote-1) The followings are p\_value matrices of two by two comparisons:

*The p\_value matrix for*

*The p\_value matrix for*

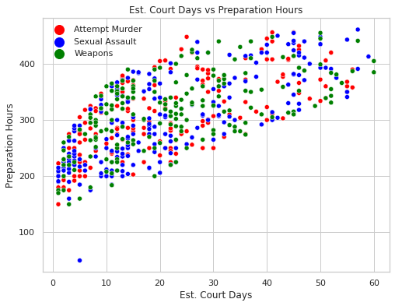
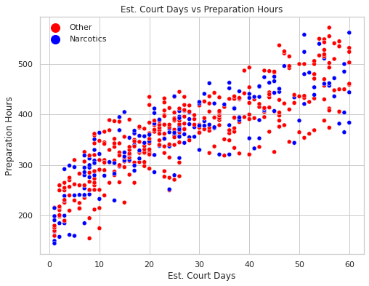
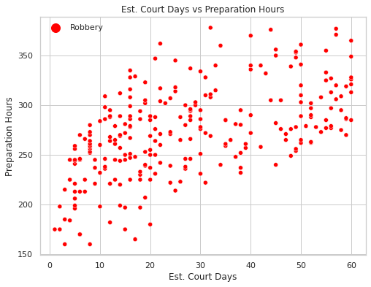
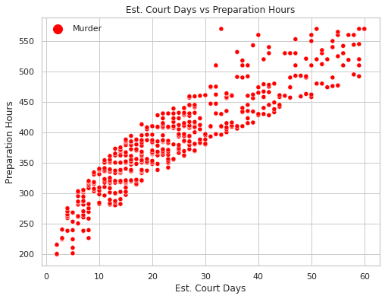


Based on α = 0.01 and the p\_value matrices we conclude:

* There is not significant statistical evidence that distribution of “Attempt Murder”, “Sexual Assault” and “Weapons” are different; and also
* there is not significant statistical evidence that distribution of “Narcotics” and “Other” are different

Therefore, we summarize the charges in the following four groups:

* *Group 1: Attempt Murder, Sexual Assault & Weapons*
* *Group 2: Murder*
* *Group 3: Narcotics and Other*
* *Group 4: Robbery*

** **

The final step before searching for the best model is ensuring that the Est. Court Days is a statistically proper predictor for the Preparation Hours. In order to do so, we need to test the correlation between Est. Court Days and Preparation Hours and verify the significant correlation between these two factors. Therefore, the test hypothesis is as follows:

and the test distribution is t-student with n-1 degrees of freedom. Using Python function pearsonr() from scipy.stat library we repeat this test for all the four groups that we created above and here are the correlation and p\_value of the tests:

|  |  |  |
| --- | --- | --- |
| **Group** | **Pearson Correlation** | **p\_value** |
| *Group 1: Attempt Murder, Sexual Assault & Weapons* | 0.69 | 0.000 |
| *Group 2: Murder* | 0.91 | 0.000 |
| *Group 3: Narcotics and Other* | 0.84 | 0.000 |
| *Group 4: Robbery* | 0.56 | 0.000 |

Since all the p\_values are almost 0.0 we reject the null hypothesis and conclude that: in all groups, there are statistical evidence that the Pearson correlation between the Est. Court Days and the Preparation Hours is not zero.

As a result, we have four groups of charges coming from four different populations and a statistically reasonable predictor “Est. Court Days”, so in the next section, we examine different potentially proper models and fit the best model based on the Regression Modeling and Least Square method by Python.

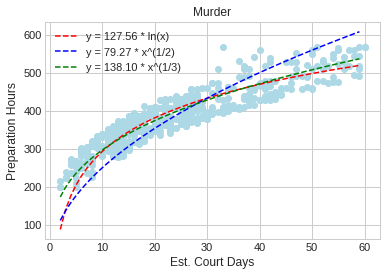
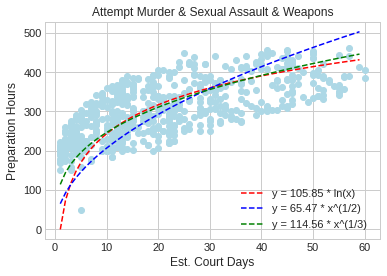
**4. Model fitting**

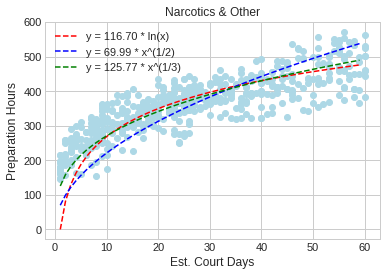
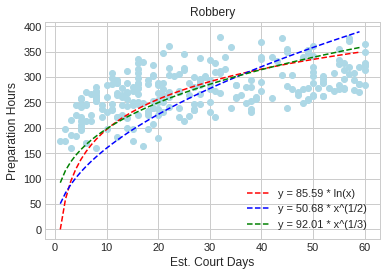
**4.1. Initial Models Comparison**

Based on the scatter plots in previous section, it seems that the Preparation Hours increases by Est. Court Days but the slop of the trend reduces gradually for the larger Est. Court Days so we observe a curved pattern in all four groups. Moreover, all the four patterns are from origin, which means that for zero Est. Court Days, Preparation Hours is reasonable zero. In another word, we will not have constant in our final model. Using Python, we fit three different models:

* Model 1:
* Model 2:
* Model 3:

By One-Way ANOVA and F-Test we statistically test each model’s fitness and by One-Sample T-Test, we examine the coefficient(s) individually for each model to ensure it is not zero.[[2]](#footnote-2)



****

|  |  |  |
| --- | --- | --- |
| **Group (Model 1)** | **One-Way ANOVA (p\_value)** | **Adjusted R-Square** |
| *Group 1: Attempt Murder, Sexual Assault & Weapons* | 0.00 | 0.943 |
| *Group 2: Murder* | 0.00 | 0.992 |
| *Group 3: Narcotics and Other* | 0.00 | 0.973 |
| *Group 4: Robbery* | 0.00 | 0.959 |

|  |  |  |
| --- | --- | --- |
| **Group (Model 2)** | **One-Way ANOVA (p\_value)** | **Adjusted R-Square** |
| *Group 1: Attempt Murder, Sexual Assault & Weapons* | 0.00 | 0.930 |
| *Group 2: Murder* | 0.00 | 0.982 |
| *Group 3: Narcotics and Other* | 0.00 | 0.965 |
| *Group 4: Robbery* | 0.00 | 0.928 |

|  |  |  |
| --- | --- | --- |
| **Group (Model 3)** | **One-Way ANOVA (p\_value)** | **Adjusted R-Square** |
| *Group 1: Attempt Murder, Sexual Assault & Weapons* | 0.00 | 0.963 |
| *Group 2: Murder* | 0.00 | 0.994 |
| *Group 3: Narcotics and Other* | 0.00 | 0.984 |
| *Group 4: Robbery* | 0.00 | 0.964 |

**4.2. Best Model selection**

Since the null hypotheses in the One-Way ANOVA tests is rejected for all three models, we use the Adjusted R-Squares to compare the model and choose the final model with the highest Adjusted R-Square. Therefore, Model 3: is selected as the best fit in all four groups:

|  |  |
| --- | --- |
| **Group (Model 3)** | Est. of Coefficient: |
| *Group 1: Attempt Murder, Sexual Assault & Weapons* | 114.5635 |
| *Group 2: Murder* | 138.1031 |
| *Group 3: Narcotics and Other* | 125.7667 |
| *Group 4: Robbery* | 92.0129 |

**4.3. Testing Key assumptions**

Since we have already fitted the best model for all groups, we need to test the key assumptions in regression modeling for this model to ensure the selected model is statistically valid. The key assumptions are as follows:

* Residuals must be Normally Distributed (Normal Distribution of Residuals)
  + Test Statistic: One Sample K – S
* Residuals must not be autocorrelated (Serial Correlation)
  + Test Statistic: Durbin-Watson
* Independent variables must not be highly intercorrelated (Multicollinearity or Variance Inflation)
  + Test Statistic: Variance Inflation Factor (VIF)

|  |  |  |  |
| --- | --- | --- | --- |
| **Group (Model 3)** | **One Sample K-S (p\_value)** | **Durbin-Waston** | **Variance Inflation Factor (VIF)** |
| *Group 1: Attempt Murder, Sexual Assault & Weapons* | 0.01 | 1.870 | 1 |
| *Group 2: Murder* | 0.02 | 1.742 | 1 |
| *Group 3: Narcotics and Other* | 0.24 | 1.606 | 1 |
| *Group 4: Robbery* | 0.21 | 1.891 | 1 |

* The p\_value of One Sample K-S tests are not less than α = 0.01 so we cannot reject the residuals normal distribution
* The Durbin-Wastons are within the range of 1.5 to 2.5 so we conclude that residuals are not autocorrelated
* We have only one independent variable in this study so we do not need to be worried for the Multicollinearity of the data (significant dependency between independent variables)

**4. Conclusion**

We fit the best models for each criminal charge, by 99% confidence level to predict the number of required preparation hours based on the estimated number of court days. The valid range of Est. Court Days (independent variable) for all charges is 1 to 60 days, since we used this range to fit the models, so models are not statistically valid outside this range. The budget management team can implement the following models for each criminal charge in an upcoming case:

* for Attempt Murder, Sexual Assault & Weapons cases:
* For Murder cases
* For Narcotics & Other cases:
* For Robbery cases:

**References**

Baker, A. J. (2010). Even when you win you lose: Targeted parents' perceptions of their

attorneys. *The American Journal of Family Therapy*, *38*(4), 292-309.

Moorhead, R. (2011). Filthy lucre: lawyers’ fees and lawyers’ ethics–what is wrong with

informed consent?. *Legal Studies*, *31*(3), 345-371

1. For T-Tests we use ttest\_ind() and for Leneve Test we use leneve() from scipy.stats library. [↑](#footnote-ref-1)
2. We use curve\_fit() from scipy.optimize library and ols() from statmodels.api library [↑](#footnote-ref-2)