Statistical Analysis on
Los Angeles Employee Payroll Analysis(2013-2016)

CSE 544: Probability and Statistics for Data Science

{swkumari,sjeevan,hagarwal,stallamraju}@cs.stonybrook.edu

https://github.com/swetakum/CSE544project

Sweta Kumari
111497926

Sagar Jeevan
111861945

Harsh Wardhan Agarwal
111465389

Sweethendra Tallamraju
111401594

1. INTRODUCTION

Hypothesis testing is an essential procedure in statistics which is used to evaluate two mutually exclusive hypothesis about a data set to determine which hypothesis is best supported by the sample data.

In a city, knowing the distribution of salary and benefits among the people gives a lot of insights to the standards of life and it also helps to understand the labour market conditions and organizational performances. Recently, after the boom of technology and data science, so many data sets were identified and open sourced for people. LA city authorities publish such payroll, crime data etc. online for public use.

We were interested to do statistical analysis on a similar data, so we chose LA payroll data of government employees ranging from 2013 to 2016. We are interested in finding interesting statistical answers and insights from the data. Majorly, we are interested in knowing below:

- Understand change and distribution of salaries over
  - years
  - quarters
  - departments
  - levels
- Find out if the annual pay can be predicted from the given dataset

The data set has a wide range of information ranging from hourly pay, annual pay, health benefits, dental benefits, overtime pay, level of the employee, department etc.

We analysed and created hypothesis along the following topics :

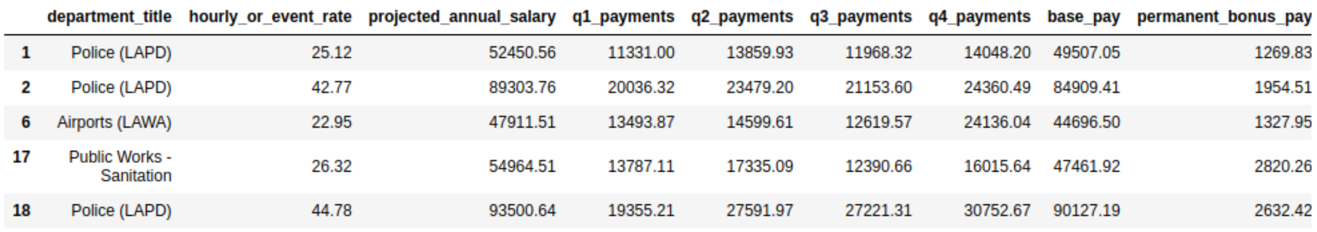
- Increase in Hourly pay and Annual Pay over the years
- Comparing salary distributions for the first half of a year and second half of the same year for a set of departments categorised as risky and non risky over the years.
- Health benefits distribution across departments categorised by seniority levels
- Predicting the projected annual salaries for the year 2016.

When we analysed the data and tested various hypothesis, we got below results:

- Annual pay and hourly pay doesn't increase over the years
- Work of non-risky departments are stagnant and is more likely to be forecasted, but the work for risky departments are very unpredictable and hence their salary distributions in the two halves of the year varies.
- Health Benefits follow same distribution over career ladders.
- Annual salaries can be predicted with very low error after required data pre-processing

2. DATASET

The dataset contains the payroll information of all Los Angeles City government employees across all departments from 2013-2016. The datset dictates various featurs associated with an employees payroll such as the salaries, benefits, department etc. Within salaries, the dataset contains several divisions such as the base_pay, hourly_rate, overtime_pay, quarterly payments (four quarters of a year). Benefits of an employee were divided as health_benefits, dental_benefits, total_benefits_cost etc. We try to leverage all these features into our analysis and try to extract the best insights from it.



	department_title	hourly_or_event_rate	projected_annual_salary	q1_payments	q2_payments	q3_payments	q4_payments	base_pay	permanent_bonus_pay
1	Police (LAPD)	25.12	52450.56	11331.00	13859.93	11968.32	14048.20	49507.05	1269.83
2	Police (LAPD)	42.77	89303.76	20036.32	23479.20	21153.60	24360.49	84909.41	1954.51
6	Airports (LAWA)	22.95	47911.51	13493.87	14599.61	12619.57	24136.04	44696.50	1327.95
17	Public Works - Sanitation	26.32	54964.51	13787.11	17335.09	12390.66	16015.64	47461.92	2820.26
18	Police (LAPD)	44.78	93500.64	19355.21	27591.97	27221.31	30752.67	90127.19	2632.42

Figure 1: Sample view of the dataset after data cleaning

2.1 Data Cleaning

In order to be able to use the dataset and get results which were statistically significant and depict close to reality results, we needed to clean the data from many dimensions first. We used the following data cleaning strategies.

- **Processing column names:** We had to process the column names of the dataset by replacing spaces with an underscore and maintain a standard naming schema so to avoid any confusion among different team members and make things easy.
- **Removing special characters:** The dataset contains lot of features that depict some sort of payment amount in dollars and hence were appended with a $ sign. There were also cases with a % sign that represented the percentage of an entity. We had to get rid of such special characters to avoid and unwanted errors in the future.
- **Typecasting numeric features to float:** In order to perform statistical analysis on any column containing numbers, it is required that all such columns are in integer/floating point format for the compiler/interpreter to recognize them as such. Thus, we had to typecast all such numeric features from strings to floating point data type.
- **Removing negative and NaN values:** There were cases where we encountered negative values at inappropriate places and we had to delete such values. Also, there were a lot of cases where the values were NaN (Not a Number). We had to process them too before performing our analysis.
- **Handling missing values:** There were several instances where certain values were missing. If they can be replaced by any statistical metric (measures of central tendancy), we replaced them else we simply delete the rows with any missing data.
- **Removing Part Time values:** There were instances where we did not want to include the employees with the job status as Part Time. For example, the health benefits of part time employees vary significantly when compared to the full time employees. They sometimes may even appear as outliers. Thus, part time employees being a very small proportion of our dataset, we remove them whenever required.
- **Removing outliers:** After plotting the box plots for several features we realized that the dataset contains a significant number of outliers which can affect the results of our analysis drastically. Thus we decided to remove outliers before performing our analysis on the dataset. We used the *Tukeys Rule for Outlier Detection* that suggests the range of (Q1-1.5*IQR, Q3+1.5*IQR) to separate out outliers.

3. PRIOR WORK

There are bunch of hypothesis tested on this data set. Following are some of hypothesis tested.

1) **Increase in Annual Salary over Years. Pay increase in year 2016. [1]**

Null Hypothesis: Annual pay does not increased in year 2016

Alternate Hypothesis: Annual pay increased in 2016

Author used Walds test and t test to reject Null Hypothesis and proved than Annual Salary increased in year 2016. Z value obtained was 7.80879 whereas t test score was 1.049946.

2) **Predicting Average Benefit Cost using Linear Regression. [1]**

Using Linear Regression Author used Annual Salary and Q1 Payments, Q2 Payments Q3 payments Q4 Payments as features to predict Average Benefit Cost. Coefficients obtained were 0.033, 0.163, 0.035, -0.36, 0.05 respectively. Coefficients suggests that it has poor correlation with target hence it was a poor model. On the other hand, we were able predict Salary effectively.

3) **Do employment benefits vary significantly between departments. [2]**

As part of initial exploration Author used plots to analyze Employment Benefits of Employees across multiple departments.Benefit plans Cost of City department is greater than DWP which inturn is greater than Police followed by Fire Department.

4. HYPOTHESES

In this section, we will discuss the various hypotheses, we tested in detail.

4.1 Topic 1 : Increase in Hourly pay and total payments over the years

In the motivation to learn whether there is any growth in terms of annual pay of employees in the city, we planned to do hypothesis testing on this. This information is useful and interesting in terms that this implicitly shows whether the city as a whole is growing financially. It also signifies the financial growth of people of the city which explicitly signifies the life index and quality of life of people of that particular city.

4.1.1 *Hypothesis:*

Increase in overall total payments and hourly pay over the years, viz. 2013-14, 2014-15

- Hourly Pay:
  H0: Hourly Pay for two years are same.
  H1: Hourly pay for two years are not same.

- Total Payments:
  H0: Total Payments for two years are same.
  H1: Total Payments for two years are not same.

List of hypothesis:

1. Hourly pay doesn't increase in year 2013-14 and 2014-15 ,using *2 sample t-test*
2. Hourly pay doesn't increase in year 2013-14 and 2014-15, using *Wald's test*
3. Total payments doesn't increase in year 2013-14 and 2014-15, using *2 sample t-test*
4. Total payments doesn't increase in year 2013-14 and 2014-15, using *Wald's test*

4.1.2 *Techniques:*

We used *Two sample T-test* and *Wald's test* to test above mentioned hypotheses.

Two sample T-test:

$$t = \frac{M_x - M_y}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}}$$

M = mean
n = number of scores per group

$$S^2 = \frac{\sum (x - M)^2}{n - 1}$$

x = individual scores
M = mean
n= number of scores in group

Assumptions : Data should be Normal

Wald's Test:

$$T = \frac{\bar{X} - \mu_o}{s/\sqrt{n}}$$

where $\bar{X}$ = Sample Mean
μ_o = True Mean
s = Sample Standard Deviation
n = size of sample

4.1.3 *Steps*

- Data Cleaning : In addition to the overall cleaning done to the dataset, we also ignored the tuples(employees) who were part time, because these may negatively impact our hypothesis testing as the part time employees have different distribution than full time employees.
- Since t-test has a pre-requisite that the underlying distribution must be normal, we plotted the data and found out that it forms a bell shaped curve.

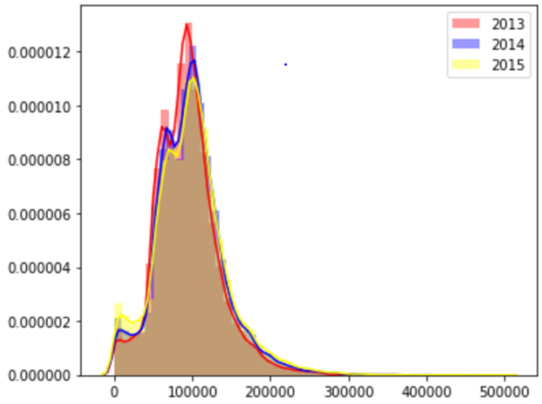
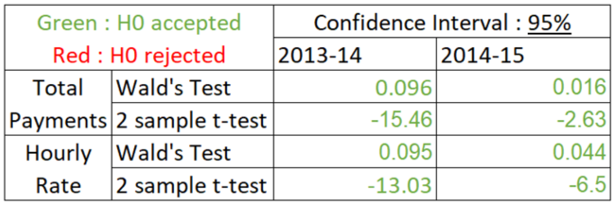


Figure 2: Normality test for Annual data
x-axis: id of data points
y-axis: data points

4.1.4 *Results*

Neither hourly pay nor total payments increases for LA government employees from year 2013 to 2014 and from 2014-2015.



Green : H0 accepted Red : H0 rejected		Confidence Interval : 95%	
		2013-14	2014-15
Total Payments	Wald's Test	0.096	0.016
	2 sample t-test	-15.46	-2.63
Hourly Rate	Wald's Test	0.095	0.044
	2 sample t-test	-13.03	-6.5

Figure 3: Statistic for Topic 1 hypotheses test

4.2 Topic 2: Department Salary Distribution Over Two Halves of A Year

Here, we consider the terms risky and non-risky as per the following justification.

Risky department: The workload that employees from this departments experience is unpredictable.

Non-risky department: The workload that employees from this departments experience is stagnant and is likely to be forecasted.

The hypothesis proposes the statement based on the assumption that pay is going to differ based on overtime hours, health benefits, and other variable pay factors.

4.2.1 Hypothesis:

H0: For risky departments, salary distribution for the first half of a year is different from the second half.

H1: For risky departments, salary distribution for the first half of a year is similar to the second half.

4.2.2 Technique Used:

- Kolmogorov-Smirnov Test for two population
- Wald's Test for two population

The tests are based on 95 percent confidence-interval.

4.2.3 Results:

The hypothesis was tested for a set of risky departments such as Police and City Attorney with non-risky departments such as City Clerk and Housing And Community.

Police (Year 2013):
Walds Test: 14.022 (Reject)
KS Test: 0.0975 with P-value = 0.017 (Reject)

Police (Year 2014):
Walds Test: 22.181 (Reject)
KS Test: 0.1482 with P-value = 0.0017 (Reject)

Police (Year 2015):
Walds Test: 9.660 (Reject)
KS Test: 0.0810 with P-value = 0.018 (Reject)

City Attorney (Year 2013):
Walds Test: 4.0964 (Reject)
KS Test: 0.2061 with P-value = 0.0688 (Reject)

City Attorney (Year 2014):
Walds Test: 2.2156 (Reject)
KS Test: 0.1751 with P-value = 0.0687 (Reject)

City Attorney (Year 2015):
Walds Test: 2.2652 (Reject)
KS Test: 0.1702 with P-value = 0.0663 (Reject)

The result speaks of risky departments having different distributions over the two halves of a year. This remains valid for all three years of 2013, 2014, and 2015 which is a clear indication that the jobs of Police and City Attorney are unpredictable. For Police department, jobs can be thought of as directly dependent on crimes and other violent acts, and that the amount has changed over the course of a particular year. The same is applied to City Attorney because crimes and other violent acts serves as a base on how these two departments workload vary.

4.2.4 Extending the hypothesis:

We can postulate one for a set of non-risky departments. We assume that these jobs are likely to be forecasted.

Hypothesis:

H0: For non-risky departments, salary distribution for the first half of a year is similar to the second half.

H1: For non-risky departments, salary distribution for the first half of a year is similar from the second half.

4.2.5 Results:

Tests below were performed for City Clerk and Housing And Community departments.

City Clerk (Year 2013):
Walds Test: 0.7506 (Accept)
KS Test: 0.1208 with P-value = 0.2016 (Accept)

City Clerk (Year 2014):
Walds Test: 0.2903 (Accept)
KS Test: 0.088 with P-value = 0.2073 (Accept)

City Clerk (Year 2015):
Walds Test: 0.1043 (Accept)
KS Test: 0.0697 with P-value = 0.2073 (Accept)

Housing And Community (Year 2013):
Walds Test: 0.7733 (Accept)
KS Test: 0.0631 with P-value = 0.080 (Accept)

Housing And Community (Year 2014):
Walds Test: 0.2044 (Accept)
KS Test: 0.0653 with P-value = 0.0819 (Accept)

Housing And Community (Year 2015):
Walds Test: 0.4578 (Accept)
KS Test: 0.0598 P-value = 0.0831 (Accept)

The results follow the hypothesis which indicate that the jobs are indeed stagnant.

4.2.6 Conclusion:

The salary distribution for risky departments vary over two halves a year. The job trends of these are unpredictable. The salary distribution for non-risky departments remain similar over two halves a year. The job trends of these are likely to be forecasted. One likely reason can be difference in overtime hours for risky departments.

4.3 Topic 3 : Health Benefits Distribution across Departments categorized by Seniority Levels

We would like to check whether health benefits follow same distribution over career ladders. For Example, Risky departments like Police and Fire Administrators follow same distribution of average health benefits across salary levels. Additionally, with in same department, different levels of employees have same distribution on health benefits.

Permutation Test and KS Test were used to test hypothesis.

For Permutation test we used 100000 permutations of all the available permutations due to limited computational capacity.

4.3.1 Hypothesis :

H0: Police Administrator I and Fire Administrator follow same distribution on average health benefit cost:

H1: Police Administrator I and Fire Administrator do not follow same distribution:

Following are the results obtained when we tried to prove this test using Permutation test and KS Test

Theme1:
Permutation test: p value is 0.36696 which is greater than 0.05(95% CI)

Theme2:
KS test: Maximum absolute difference of two CDFs is 0.06 whereas threshold is 0.652.

Theme3:

Police Administrator II and Fire Administrator follow same distribution on average health benefit cost:

Permutation test: p value is 0.36785 which is greater than 0.05

4.3.2 Results :

Following are some of the additional inferences we derived based on given data.

- **Police Administrator II and Fire Administrator follow same distribution on average health benefit cost:**

  KS test: maximum absolute difference of two CDFs is 0.066 whereas threshold is 0.6767.

- **Police Detective I and Police Detective II follow same distribution on average health benefit cost:**

  Permutation test: p value is 0.36696 which is greater than 0.05

  KS test: maximum absolute difference of two CDFs is 0.006065 whereas threshold is 0.03425257

- **Police Detective I and Police Detective III follow same distribution on average health benefit cost:**

  Permutation test: p value is 0.3671 which is greater than 0.05

- **Police Detective II and Police Detective III follow same distribution on average health benefit cost:**

  Permutation test: p value is 0.36737 which is greater than 0.05

- **Police Officer I and Police Officer II follow same distribution on average health benefit cost:**

  Permutation test: p value is 0.36437 which is greater than 0.05

- **Police Officer I and Police Officer III follow same distribution on average health benefit cost:**

  Permutation test: p value is 0.36588 which is greater than 0.05

- **Police Officer II and Police Detective III follow same distribution on average health benefit cost:**

  Permutation test: p value is 0.36494 which is greater than 0.05

- **Fire Inspector I and Fire Inspector II follow same distribution on average health benefit cost:**

  Permutation test: p value is 0.3678 which is greater than 0.05

- **Financial Manager I and Financial Manager II have same distribution on average health benefit cost:**

  KS test: maximum absolute difference of two CDFs is 0.04212 whereas threshold is 0.399

- **Accountant I and Accountant II follow same distribution on average health benefit cost:**

  Permutation test: p value is 0.36778 which is greater than 0.05

4.3.3 Conclusion

We can conclude that Average Health Benefits follow same distributions over career Ladders.

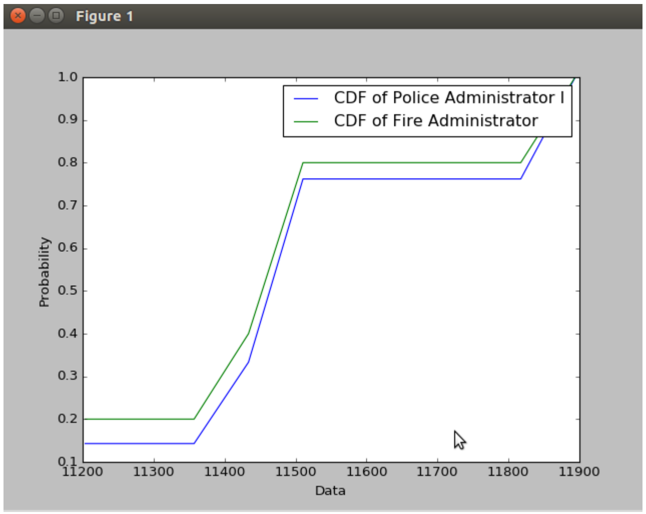
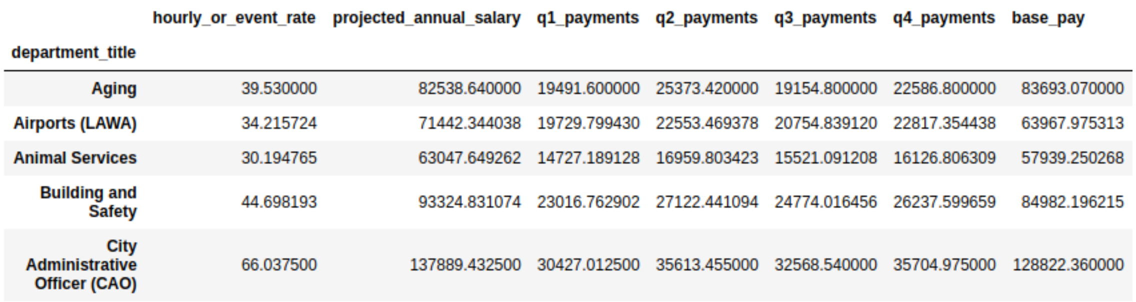


Figure 4: Comparing CDF's of Police AdministratorI and Fire Administrator

4.4 Topic 4 : Predicting the projected annual salaries for the year 2016

Every organization/company releases an annual salary statement for each of its employees to have a rough estimate of the pay expenses and plan the budget for themselves for the coming year. In such cases, it is required to have good estimators for the prediction else it may lead to wrong estimation of the fore coming budget. With this motivation, we planned to check if we can come up with a good estimator for the projected_annual_salary column. For our analysis,



	hourly_or_event_rate	projected_annual_salary	q1_payments	q2_payments	q3_payments	q4_payments	base_pay
department_title							
Aging	39.530000	82538.640000	19491.600000	25373.420000	19154.800000	22586.800000	83693.070000
Airports (LAWA)	34.215724	71442.344038	19729.799430	22553.469378	20754.839120	22817.354438	63967.975313
Animal Services	30.194765	63047.649262	14727.189128	16959.803423	15521.091208	16126.806309	57939.250268
Building and Safety	44.698193	93324.831074	23016.762902	27122.441094	24774.016456	26237.599659	84982.196215
City Administrative Officer (CAO)	66.037500	137889.432500	30427.012500	35613.455000	32568.540000	35704.975000	128822.360000

Figure 5: Dataset grouped by departments to computed estimated q1, q2,q3 and q4 payments

we plan to use Linear Regression to come up with our predictions. We plan to train the model on the data from the years 2013-2015 and predict for the year 2016. The hypothesis for our problem statement can be stated as below.

4.4.1 Hypothesis:

H0: We can predict projected annual salary with high accuracy
H1: We cannot predict projected annual salary with high accuracy
Technique used: Linear regression with variations in the training data.

4.4.2 Steps:

- Subset dataset with year less than equal to 2015 for training (X_train) and year equal to 2016 for testing (y_train) purpose.
- Keep only important features and discard the rest.
- Apply data pre-processing to ensure data validity. (-ve values, invalid zeros, NaN etc)
- Remove outliers
- Fit the data using Linear Regression model. (We used sklearn)
- Prepare X_test data by mean estimates of historical data and predict using the generated model.
- Check the results against the y_test and compute SSE and MAPE

4.4.3 Regression Equation: (Generic form)

Y = B0 + B1(hourly_rate) + B2(q1_pay) + B3(q2_pay) + B4(q3_pay) + B5(q4_pay) + B6(base_pay) + B7(permanent_bonus)

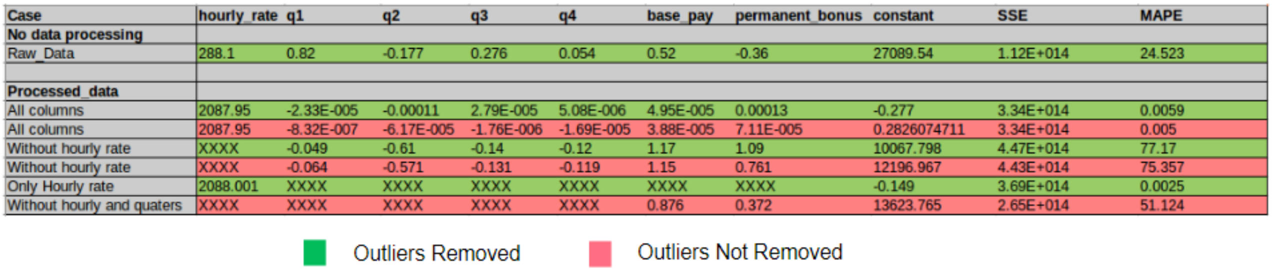
Based on the correlations between all the features of the dataset, we came up with seven important features that we could use for our predictions. However, we faced a challenge in finding the appropriate data. For instance, imagine that we are on a time stamp at the beginning of the financial year of 2016 and we need to make predictions for the upcoming years annual salaries of all the employees. At this point of time we will only have the information about the hourly_rate, base_pay and permanent_bonus of the employees as it is fixed. Whereas, the quarterly payments data will be available at the end of every quarter. Thus, we will need to come up with an estimator for the quarterly payments that can estimate the values for these features before time.

In our case, we grouped the entire dataset (after data processing) by department and for each department, we take the mean of the respective quarterly values. While predicting the annual salary for a particular employee, we find the department to which the employee belongs to and plug in the estimated mean values and substitute them in the regression equation as the quarterly payment details for that employee and generate the prediction (Figure 3).

4.4.4 Results

We have tabulated the results in a table shown in the Figure 4. The rows show the various scenarios under which we run the Linear Regression and the columns show the regression coefficients of the respective feature listed on each column heading. We performed the analysis under various scenarios to understand the way statistics of a model change under certain conditions. We performed regression analysis before and after data pre-processing just to see the effect of processing the data. We also performed regression analysis on data with outliers and data without the outliers to understand the effect of removing outliers from our dataset. We can observe from the table that the MAPE value drastically dropped from running linear regression on data before data processing to running linear regression on data after data processing. Which means our data processing helped us in improving our predications in a positive way. Now, we notice that the MAPE value is extremely small (0.0059) which makes our predictions extremely accurate. Such high accuracy is rare to achieve.

Additionally, by computing the correlation matrices (Figure 5), we also find that the coefficient of the hourly_rate feature is significantly high when compared to the coefficients of other features. This means that hourly_rate is the most important feature among other features to have such a high weightage. Investigating this issue further and to establish that the above finding about hourly_rate to be true, we plan to train the regression model without the hourly_rate feature and check the results. The results turn out to be very poor on predictions thus establishing that hourly_rate is indeed a very important feature for prediction.



Case	hourly_rate	q1	q2	q3	q4	base_pay	permanent_bonus	constant	SSE	MAPE
No data processing										
Raw_Data	288.1	0.82	-0.177	0.276	0.054	0.52	-0.36	27089.54	1.12E+014	24.523
Processed_data										
All columns	2087.95	-2.33E-005	-0.00011	2.79E-005	5.08E-006	4.95E-005	0.00013	-0.277	3.34E+014	0.0059
All columns	2087.95	-8.32E-007	-6.17E-005	-1.76E-006	-1.69E-005	3.88E-005	7.11E-005	0.2826074711	3.34E+014	0.005
Without hourly rate	XXXX	-0.049	-0.61	-0.14	-0.12	1.17	1.09	10067.798	4.47E+014	77.17
Without hourly rate	XXXX	-0.064	-0.571	-0.131	-0.119	1.15	0.761	12196.967	4.43E+014	75.357
Only Hourly rate	2088.001	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	-0.149	3.69E+014	0.0025
Without hourly and quaters	XXXX	XXXX	XXXX	XXXX	XXXX	0.876	0.372	13623.765	2.65E+014	51.124

🟩 Outliers Removed 🟥 Outliers Not Removed

Figure 6: Results obtained for different regression models

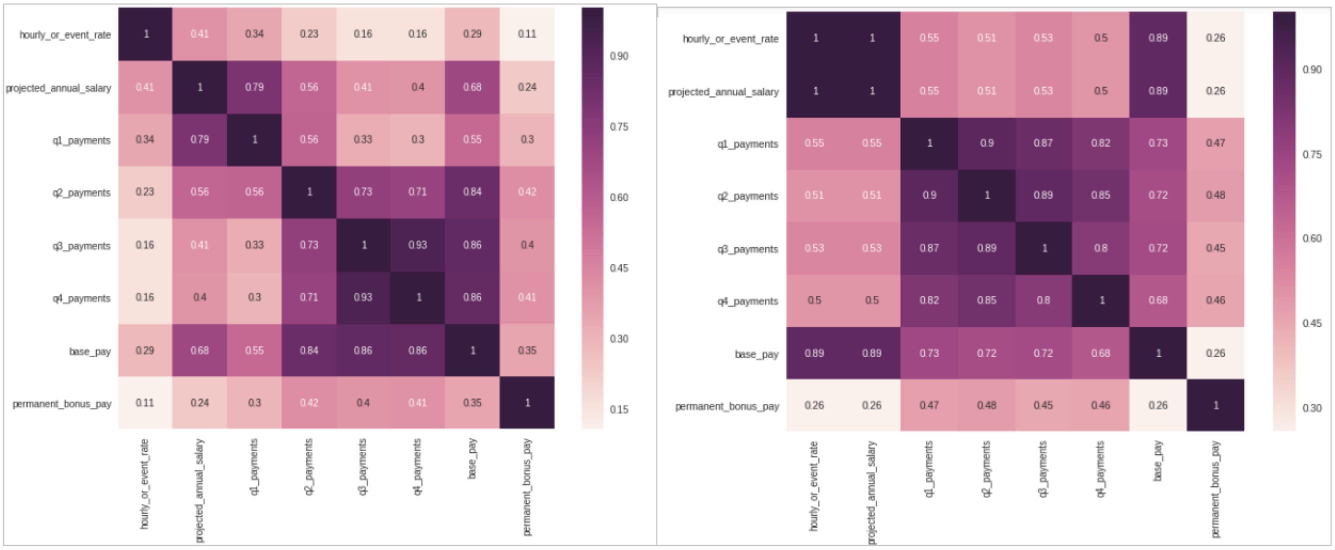


Figure 7: Correlation matrices before (left) and after (right) data processing

We also train a regression model using only hourly_rate as our input and the MAPE turned out to be the lowest of all our analysis (0.0025). However, we were skeptic about such low MAPE value which is hard to achieve in real world analysis. Hence we looked at the correlations between the important features used before and after data pre-processing. We observed that the correlation between our target variable projected_annual_salary and hourly_rate was 0.41 before data processing. But after the data pre-processing, the correlation between the two features went straight up to 1 which is the sole reason for having such low MAPE value and accurate predictions (depicted by Figure 5).

4.4.5 *Conclusion*

The projected annual salary can be predicted with high probability if there are few attributes that correlate well with the target variable. Our data pre-processing did a real good job resulting in high correlations between the predicting features and the target variable. It led us to conclude that the projected_annual_salary feature in the dataset has a very high correlation with the hourly rate of the employees (high correlation than the base_pay itself) and can be predicted with high accuracy using just the hourly_rate feature.

5. FUTURE WORK

1) **Predicting the rise or downfall of a department.**

The total payment over the years can be performed on department levels. Based on the distributions obtained, we can check if there has been an increase or decrease of total payments over the years. This can lead to interpretations such as revenue generated by the department and can help make serious conclusion on whether it will cease to exist or endure.

2) **The dataset can be coupled with other related datasets.**

To draw a better inference on our results, multiple datasets can be coupled. For example, the police department data can be coupled with a Los Angeles crime dataset over the same years, to get a better understanding on the dependency of police department (variation of number of employees) on crime-rates.

3) **Various other benefits distributed across departments categorized by seniority levels.**

Other benefits of employees such hourly pay, permanent bonus pay, longevity_bonus_pay can be used to learn the distributions and to draw inferences.

4) **Taking into account, the part-timers.**

The results were performed only for full-timers. A better understanding of a department can be achieved by including part-timers. Department level analysis of part-timers along with full-timers help us delineate the differences on how each department receives them.

6. REFERENCES

[1] Kaggle LA Payroll Challenge https://www.kaggle.com/bharath25/hypothesis-testing-and-predictive-analysis?scriptVersionId=876226

[2] Kaggle initial comments https://www.kaggle.com/annawolak/initial-exploration/comments

[3] LA City Payroll Data https://data.lacity.org/browse?category=Payrollutf8=%E2%9C%93

[4] https://data.lacity.org/Payroll/City-Employee-Payroll/pazn-qyym

[5] https://www.kaggle.com/cityofLA/city-payroll-data