Math and Statistics Project Report

Muhammad Talha 20I-2212

National University of Computer and Emerging Sciences, Islamabad

Contents

[Preprocessing 2](#_Toc60610646)

[Missing values 2](#_Toc60610647)

[Null Values 2](#_Toc60610648)

[Removing $ sign 3](#_Toc60610649)

[Filling Missing Values 4](#_Toc60610650)

[Task 1 5](#_Toc60610651)

[Null Hypothesis: 5](#_Toc60610652)

[Alternative Hypothesis: 5](#_Toc60610653)

[Task 2 6](#_Toc60610654)

[Null Hypothesis 6](#_Toc60610655)

[Alternative Hypothesis 6](#_Toc60610656)

[Task 3 7](#_Toc60610657)

[Null Hypothesis 7](#_Toc60610658)

[Alternative Hypothesis 7](#_Toc60610659)

[Task 4 8](#_Toc60610660)

[Null Hypothesis 8](#_Toc60610661)

[Alternative Hypothesis 8](#_Toc60610662)

[Task 5 9](#_Toc60610663)

[Null Hypothesis 9](#_Toc60610664)

[Alternative Hypothesis 9](#_Toc60610665)

[Task 6 10](#_Toc60610666)

[Null Hypothesis 10](#_Toc60610667)

[There is relationship between Department and Employment type 10](#_Toc60610668)

[Alternative Hypothesis 10](#_Toc60610669)

[Task 7 12](#_Toc60610670)

[Null Hypothesis 12](#_Toc60610671)

[There is relationship between Department and Employment type 12](#_Toc60610672)

[Alternative Hypothesis 12](#_Toc60610673)

[Task 8 12](#_Toc60610674)

[Null Hypothesis 12](#_Toc60610675)

[Other than Water and Power (DWP) pay more benefit to employees 12](#_Toc60610676)

[Alternative Hypothesis 12](#_Toc60610677)

[More result: 13](#_Toc60610678)

[Task 9 13](#_Toc60610679)

[Null Hypothesis 13](#_Toc60610680)

[employees of Recreation and Parks Department have denied of their Longevity Bonus Pay 13](#_Toc60610681)

[Alternative Hypothesis 13](#_Toc60610682)

[Task 10 13](#_Toc60610683)

[Null Hypothesis 13](#_Toc60610684)

[Harbor have more average health cost than water and power dept 13](#_Toc60610685)

[Alternative Hypothesis 13](#_Toc60610686)

[Figure 1 loading data with applying NA in missing values 4](#_Toc60610687)

[Figure 2 NA Value Counter Function 4](#_Toc60610688)

[Figure 3 Missing values Graph 5](#_Toc60610689)

[Figure 4 Snippet of columns with sign 5](#_Toc60610690)

[Figure 5 Removing $ Sign from Variables Entries 6](#_Toc60610691)

[Figure 6 Dataset After Removing $ Signs 6](#_Toc60610692)

[Figure 7 Impute Data Command 6](#_Toc60610693)

[Figure 8 Filling Categorical NA Values 7](#_Toc60610694)

[Figure 9 Clean Dataset 7](#_Toc60610695)

[Figure 10 Commands Task 1 8](#_Toc60610696)

[Figure 11 Task 1 Hypothesis Result 8](#_Toc60610697)

[Figure 12 Task 2 Hypothesis Commands 9](#_Toc60610698)

[Figure 13 Task 2 Hypothesis Results 9](#_Toc60610699)

[Figure 14 Task 3 Hypothesis Commands 10](#_Toc60610700)

[Figure 15 Task 3 Hypothesis Results 10](#_Toc60610701)

[Figure 16 Task 4 Hypothesis Commands 11](#_Toc60610702)

[Figure 17 Task 4 Hypothesis Results 11](#_Toc60610703)

[Figure 18 Task 5 Hypothesis Commands 12](#_Toc60610704)

[Figure 19 Task 5 Hypothesis Results 12](#_Toc60610705)

[Figure 20 Task 5 Outliers 12](#_Toc60610706)

[Figure 21 Task 6 Hypothesis Commands 13](#_Toc60610707)

[Figure 22 Task 6 Hypothesis Results 13](#_Toc60610708)

[Figure 23 Task 6 Scatter Plot 13](#_Toc60610709)

[Figure 24 Task 7 Hypothesis Results 14](#_Toc60610710)

[Figure 25 Task 8 Hypothesis Results 14](#_Toc60610711)

[Figure 26Task 9 results 15](#_Toc60610712)

[Figure 27Task 10 Results 16](#_Toc60610713)

# Preprocessing

Many real-world datasets contain missing/NA values for various reasons. They are often encoded as NA, blanks, or any other placeholders. Training a model with a dataset that has a lot of missing values can drastically impact the machine learning model’s quality. As told in the project that the dataset contains missing and null values. Therefore, operations have been applied to the data to make the data more reliable and satisfy the project requirement.

## Missing values

In the original dataset missing values were not showing as “NA” therefore while loading the data from the file [fill = T, na.strings = "" ] has been added as shown in figure 1.

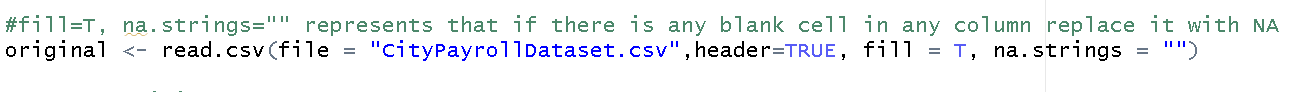


Figure loading data with applying NA in missing values

## Null Values

To count the null values in numbers a custom function has been made, this can also be achieved by sapply() command.

A picture containing text

Description automatically generated

Figure NA Value Counter Function

It can be observed that some columns have NA values up to 35% of the total number. These null values can be problematic in the analysis of the dataset. Therefore these have to be counted according to the proper method.

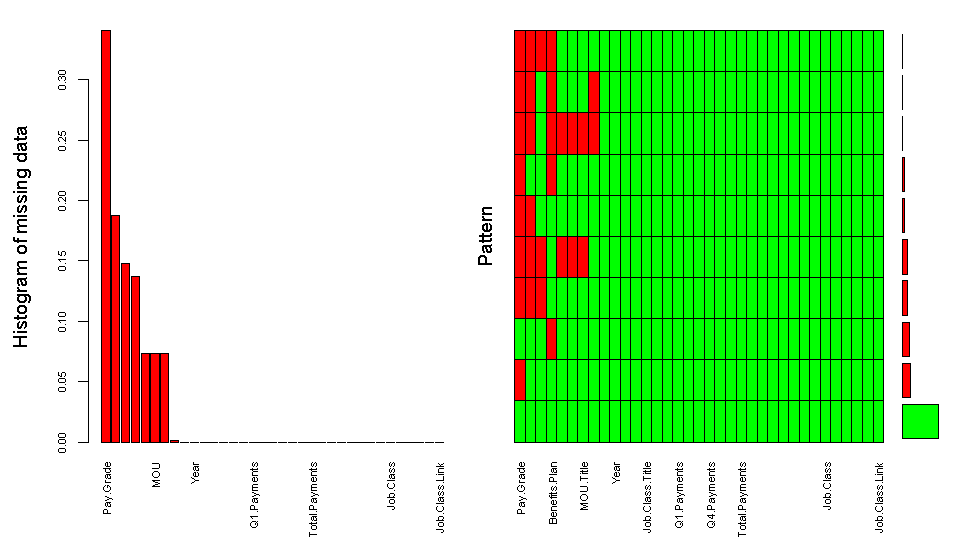


Figure Missing values Graph

## Removing $ sign

It has been observed that the numeric columns have $ and % sign as shown in figure 5. To perform arithmetic operations on columns they must be converted with no sign.

Table

Description automatically generated

Figure Snippet of columns with sign

To perform this operations all of the numeric columns are stored in array and in a loop a command performed operation to remove the signs as shown in Figure 5.

A picture containing text

Description automatically generated

Figure Removing $ Sign from Variables Entries

After removing the signs the dataset looks as shown in Figure 6.

Table

Description automatically generated

Figure Dataset After Removing $ Signs

## Filling Missing Values

To fill the NA values in the dataset many techniques have been tested which include decision tree, logistic regression, linear modeling, and mean of the dataset. All of these techniques can be applied by changing a single parameter method with the respective technique keyword from the documentation. For instance, “ppm” is used to impute data with a mean value.

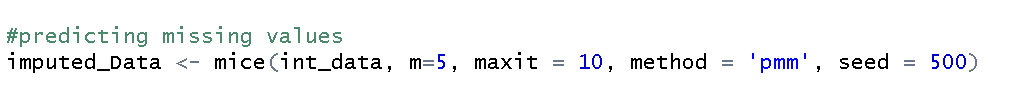


Figure Impute Data Command

This method finds the mean values for all the columns in the int\_data dataset and stores their results in the imputed\_data object. After that, these values are inserted into the dataset one by one. After filling the dataset with new values new dataset has no null values in numerical variables. As shown in the above graph, some variables contain missing values. These null values are filled with the mod of their present values. As shown in Figure 8.

Text

Description automatically generated

Figure Filling Categorical NA Values

This method is performed on all the categorical variables. In the end, the dataset contains no null/NA value it. This can be observed visually in Figure 9.

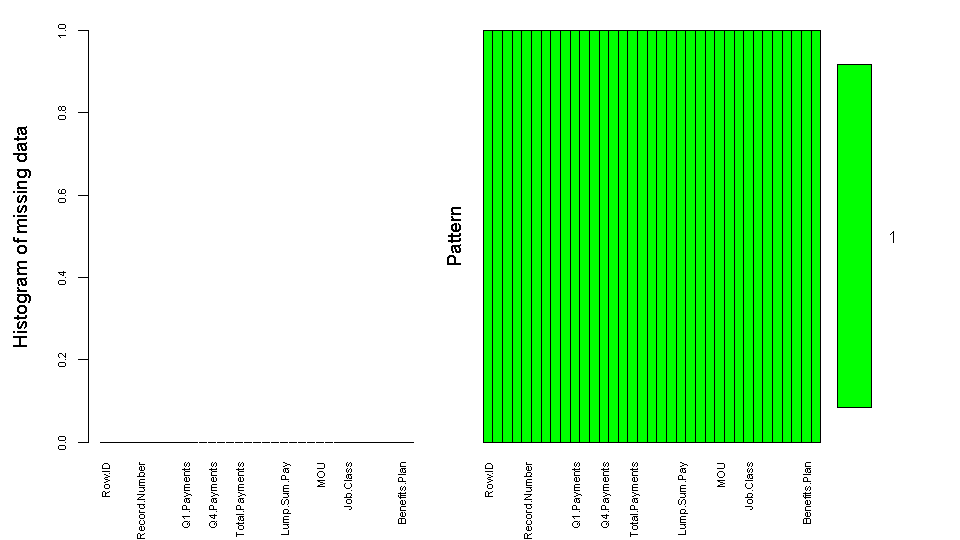


Figure Clean Dataset

# Task 1

## Null Hypothesis:

Employees who get permanent bonus are "likely" to be from police officers-II

## Alternative Hypothesis:

Employees who get permanent bonus are "not likely" to be from police officers-II

Graphical user interface, text, application, email

Description automatically generated

Figure Commands Task 1

From the above figure it can be seen that in null hypothesis police officers II are getting temporary bonus whether in the alternative hypothesis, a sample has been drawn from the clean dataset. The sample is of size 40 because it is big dataset and sample can be skewed(less than 120) and it is randomly select (replace=T). This satisfies the conditions studied in class.

Text

Description automatically generated

Figure Task 1 Hypothesis Result

After running the command t.test() all the information to validate hypothesis have been obtained. From the test output, it can be seen that the p-values is greater than 0.05 and mean of the sample is between 95% CI values. Therefore, alternative hypothesis Employees who get permanent bonus are "not likely" to be from police officers-II have been rejected.

# Task 2

## Null Hypothesis

Employees who get permanent bonus are "likely" to be from Public Works - Sanitation

## Alternative Hypothesis

Employees who get permanent bonus are "not likely" to be from Public Works – Sanitation

Graphical user interface, text, application, email

Description automatically generated

Figure Task 2 Hypothesis Commands

From the above figure it can be seen that in null hypothesis Public Works - Sanitation are getting permanent bonus whether in the alternative hypothesis, employees who get permanent bonus are not likely from Public Works – Sanitation. A sample has been drawn from the clean dataset. The sample is of size 40 because it is big dataset and sample can be skewed(less than 120) and it is randomly select (replace=T). This satisfies the conditions studied in class.

Text, letter

Description automatically generated

Figure Task 2 Hypothesis Results

After running the command t.test() all the information to validate hypothesis have been obtained. From the test output, it can be seen that the p-values is greater than 0.05 and mean of the sample is between 95% CI values. Therefore, alternative hypothesis employees who get permanent bonus are not likely from Public Works – Sanitation have been rejected.

## Task 3

## Null Hypothesis

Employees working in Water and Power (DWP) Department “likely” being employed overtime

## Alternative Hypothesis

Employees working in Water and Power (DWP) Department “not likely” being employed overtime

Graphical user interface, text, application, email

Description automatically generated

Figure Task 3 Hypothesis Commands

From the above figure it can be seen that in null hypothesis Employees working in Water and Power (DWP) Department “likely” being employed overtime whether in the alternative hypothesis, Employees working in Water and Power (DWP) Department “not likely” being employed overtime. A sample has been drawn from the clean dataset. The sample is of size 40 because it is big dataset and sample can be skewed(less than 120) and it is randomly select (replace=T). This satisfies the conditions studied in class.

Text, letter

Description automatically generated

Figure Task 3 Hypothesis Results

After running the command t.test() all the information to validate hypothesis have been obtained. From the test output, it can be seen that the p-values is greater than 0.05 and mean of the sample is between 95% CI values. Therefore, alternative hypothesis employees working in Water and Power (DWP) Department “not likely” being employed overtime have been rejected.

## Task 4

## Null Hypothesis

Employees who work Part Time likely to be from the Airports (LAWA) Department

## Alternative Hypothesis

Employees who work Part Time not likely to be from the Airports (LAWA) Department

Graphical user interface, text, application, email

Description automatically generated

Figure Task 4 Hypothesis Commands

From the above figure it can be seen that in null hypothesis Employees who work Part Time likely to be from the Airports (LAWA) Department whether in the alternative hypothesis, Employees who work Part Time are not likely to be from the Airports (LAWA) Department. A sample has been drawn from the clean dataset. The sample is of size 40 because it is big dataset and sample can be skewed(less than 120) and it is randomly select (replace=T). This satisfies the conditions studied in class.

Text, letter

Description automatically generated

Figure Task 4 Hypothesis Results

The under consideration columns are categorical therefore chisq.test() run on the variables to obtain all the information to validate the alternative hypothesis. From the test output, it can be seen that the p-values is much smaller than 0.05 and mean of the sample is not in between 95% CI values. Therefore, alternative hypothesis Employees who work Part Time not likely to be from the Airports (LAWA) Department is true and accepted.

## Task 5

## Null Hypothesis

Police (LAPD) Department has experienced the highest pay raise

## Alternative Hypothesis

Police (LAPD) Department has not experienced the highest pay raise

Graphical user interface, text, application, email

Description automatically generated

Figure Task 5 Hypothesis Commands

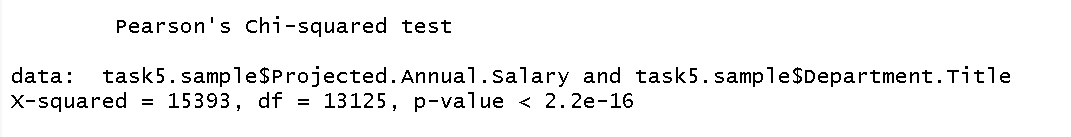


Figure Task 5 Hypothesis Results

It can be seen that alternative hypothesis Police (LAPD) Department has not experienced the highest pay raise is true.

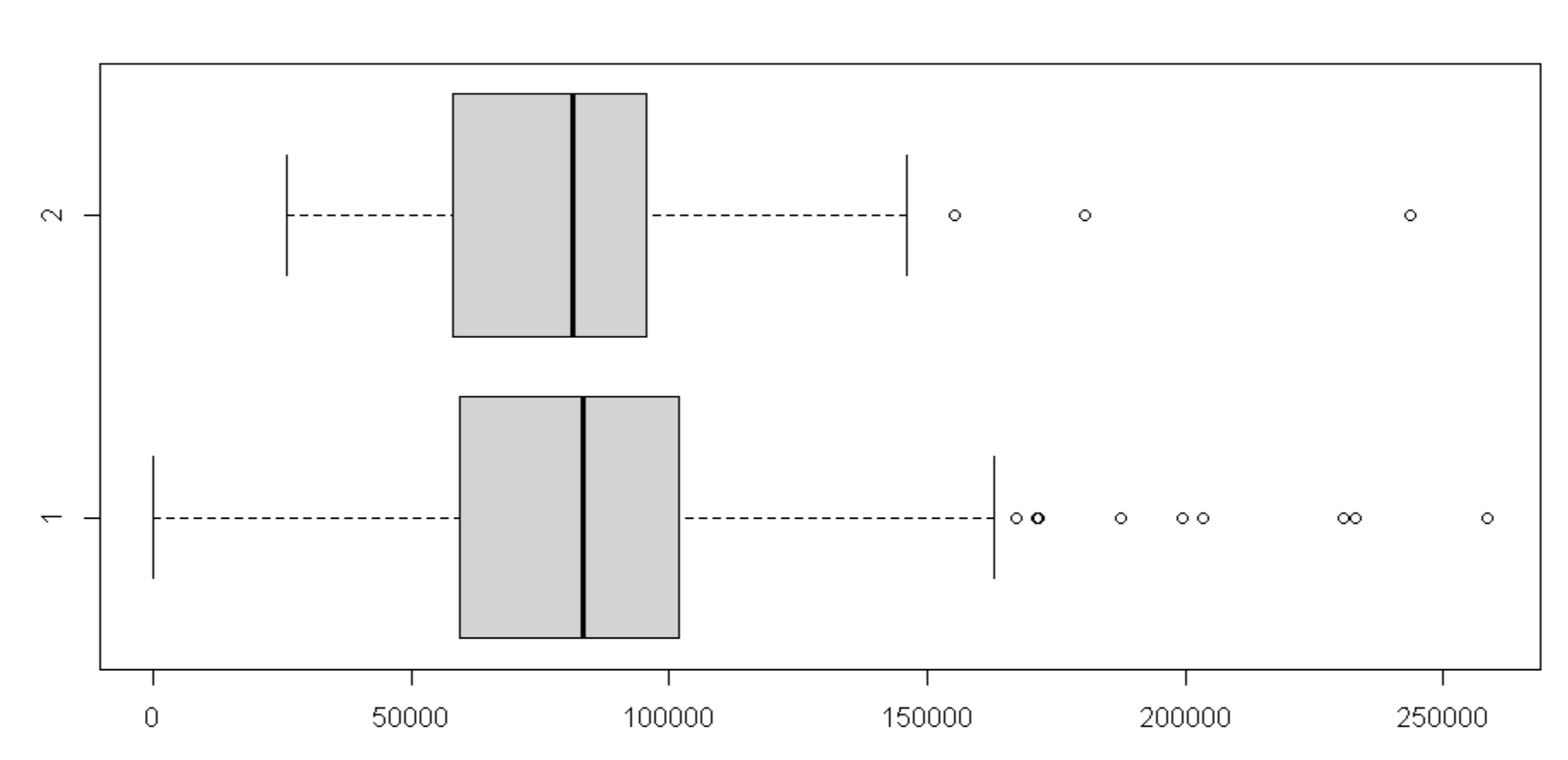


Figure Task 5 Outliers

Although this sample have many outliers, results may vary if outliers are removed.

## Task 6

## Null Hypothesis

## There is relationship between Department and Employment type

## Alternative Hypothesis

There is “no relationship” between Department and Employment type

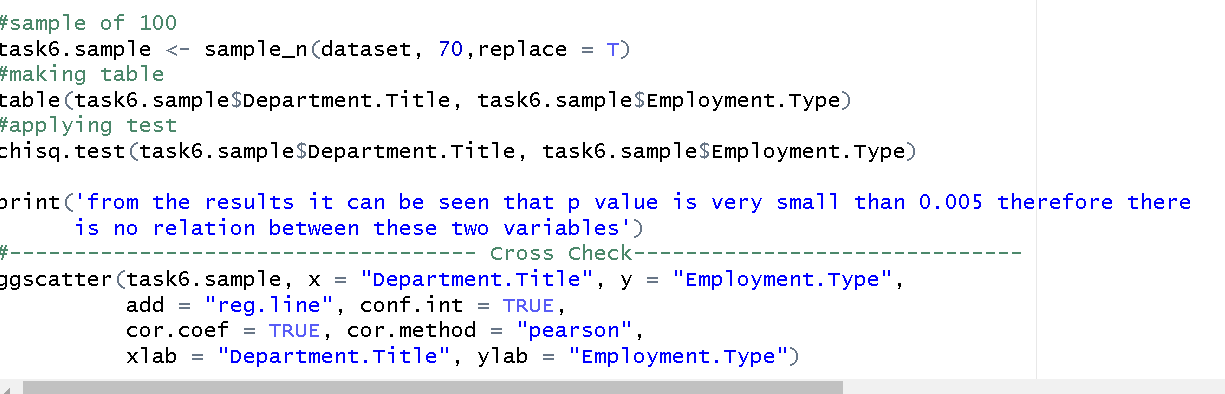


Figure Task 6 Hypothesis Commands

The above commands have been run to test the null and alternative hypothesis,

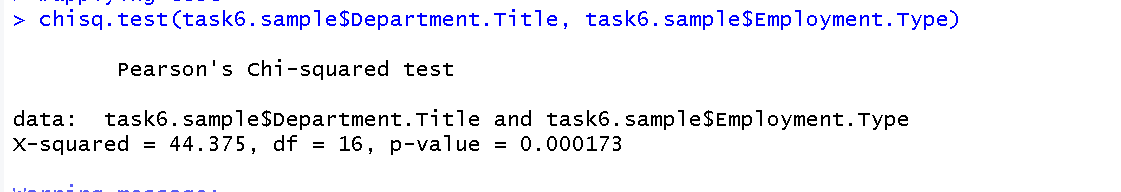


Figure Task 6 Hypothesis Results

This shows that the p value is less than .005 therefore, there is no relation between employment type and department title.

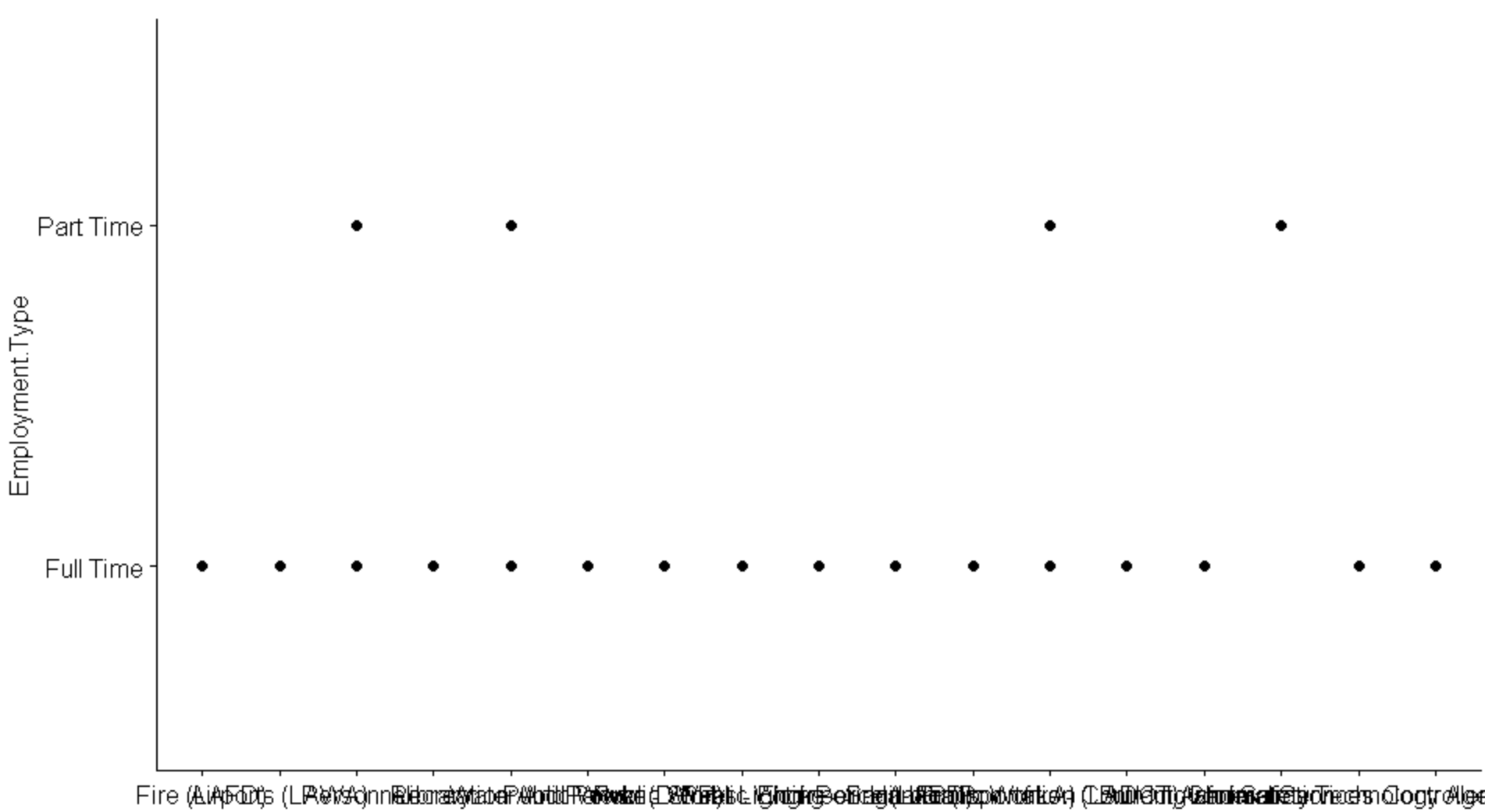


Figure Task 6 Scatter Plot

This also shows that there is no relation between the above two variables.

## Task 7

## Null Hypothesis

## There is relationship between Department and Employment type

## Alternative Hypothesis

There is “no relationship” between Department and Employment type

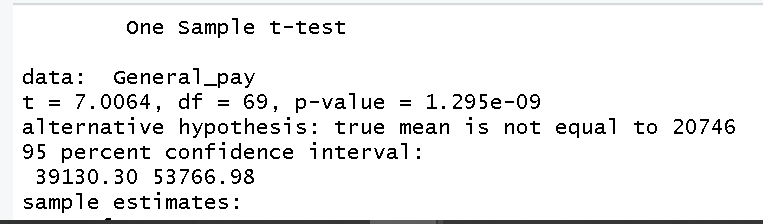


Figure Task 7 Hypothesis Results

it has been observed that p is less than 0.005 therefore alternative hypothesis Average Base Pay of employees in General Services Department is not more than $20,746 therefore alternative hypothesis is accepted.

## Task 8

## Null Hypothesis

## Other than Water and Power (DWP) pay more benefit to employees

## Alternative Hypothesis

Water and Power (DWP) Department pays on average more Benefit cost to its employees than all other departments.

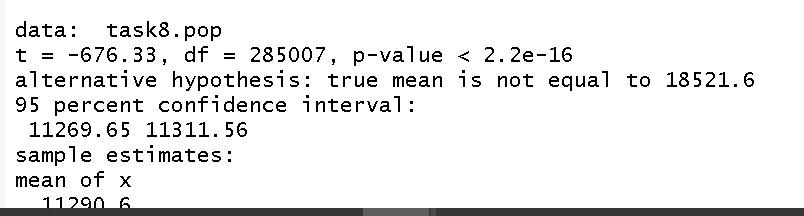
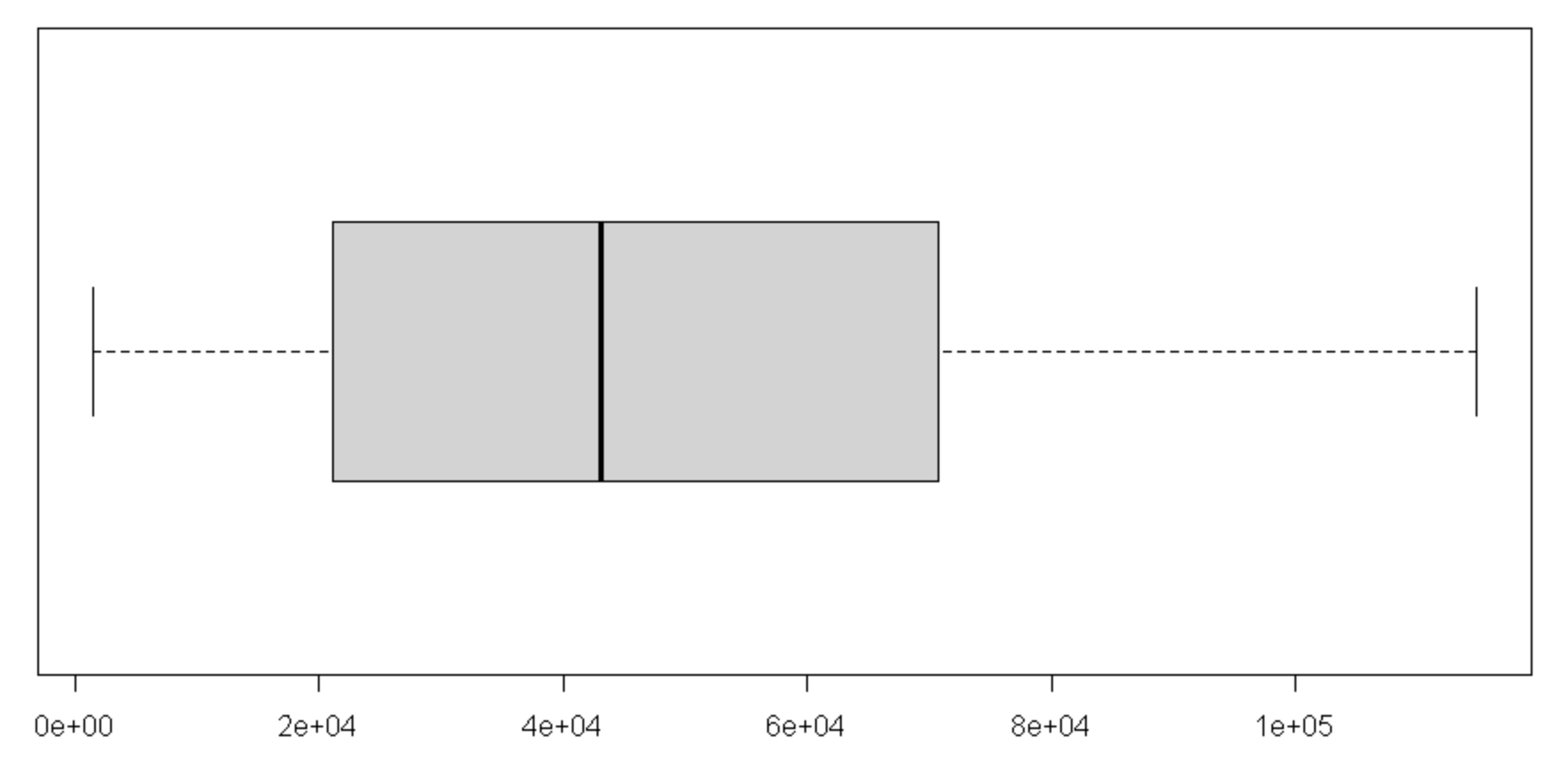


Figure Task 8 Hypothesis Results

it has been observed that p is less than 0.005 therefore alternative hypothesis Water and Power (DWP) Department pays on average more Benefit Cost to its employees than all other departments is accepted

## More result:



## Task 9

## Null Hypothesis

## employees of Recreation and Parks Department have denied of their Longevity Bonus Pay

## Alternative Hypothesis

employees of Recreation and Parks Department haven't denied of their Longevity Bonus Pay

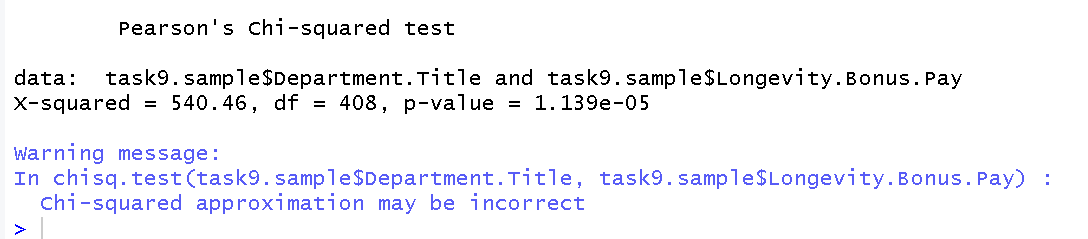


Figure Task 9 results

Alternative hypothesis have been accepted

## Task 10

## Null Hypothesis

## Harbor have more average health cost than water and power dept

## Alternative Hypothesis

Water Dept has more average health cost than Harbor dept.

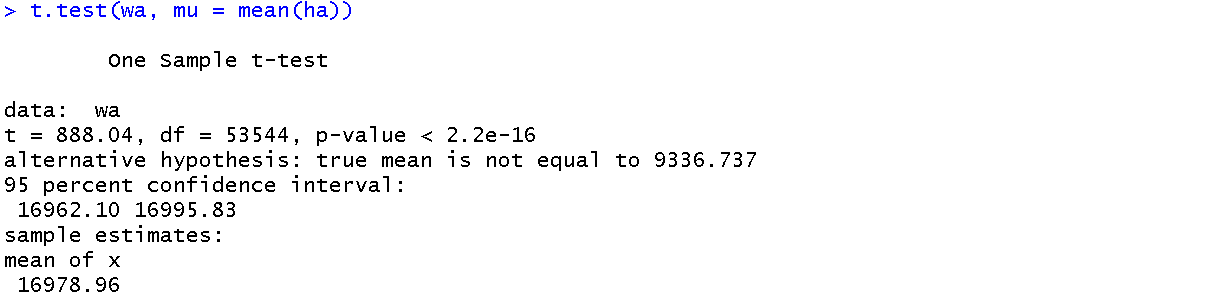


Figure Task 10 Results

Alternative hypothesis has been accepted