**MID-TERM PROJECT REPORT**

**TITLE OF THE PROJECT:** San Francisco\_Salary\_3Years. (Salaries of San Francisco residents for 3 years)

**DATA SET SOURCE:**

[**https://www.kaggle.com/datasets/bf0d8d6198ad19892386db65f867cc2ff9f4ac23bd4bb5decc3ccae8aca51a2b?resource=download**](https://www.kaggle.com/datasets/bf0d8d6198ad19892386db65f867cc2ff9f4ac23bd4bb5decc3ccae8aca51a2b?resource=download)

****

**DESCRIPTION ABOUT THE DATA:**

The data chosen consists of fields like

|  |  |
| --- | --- |
| **ID** | **Benefits** |
| **EmployeeName** | **TotalPay** |
| **JobTitle** | **TotalPayBenefits** |
| **BasePay** | **Year** |
| **OvertimePay** | **Department** |
| **OtherPay** |  |

As the dataset is huge, I have filtered top 100 rows of the data of academic year 2013 based on the total pay of the Employee. I am attaching the Pdf of the filtered data. I will Use this data to apply the concepts I learnt in the consecutive modules and perform the analysis.

**Applying lessons learnt from Module 1:**

1. **Calculate the Mean and Standard Deviation of the Total Pay for unique job role and will round the answer to 4 decimal places. (Assume the job role to be Carpenter)**

Here am directly taking the entire data set directly from the Kaggle to find out the total pay for Job role: Carpenter.

|  |  |
| --- | --- |
| Mean | 74362.0100 |
| Standard Deviation | 24304.5800 |
| Median | 83504.4100 |
| Mode | 1563.0000 |
| Standard Error | 1620.3050 |
| Sample Variance | 593349657.4000 |
| Kurtosis | 1.9966 |
| Skewness | -1.7298 |
| Minimum | 0 |
| Maximum | 108118.7000 |
| Range | 108118.7000 |
| Sum | 16731452.49 |
| Count | 226 |

**Out of the 226 observations of the Carpenter Job role, The mean salary of the Carpenter Job role is 74362.0100 and the Standard Deviation of the data filtered is 24304.5800.( Rounded the values upto 4 decimal places.)**

**(NOTE: The calculations are done using excel.)**

**Applying lessons learnt from Module 2:**

1. **Calculate the mean and Standard Deviation of the salaries ranging from $2,00,000 to $4,00,000 and construct 95% Confidence interval and round off the calculated value to 2.**

Well, we will calculate mean, standard deviation and construct the confidence interval taking into account the data that was sorted based on the Totalpay between 200k-400k and the academic year to be 2013, which we already listed top 100 out of that. Now we will find the values considering the filtered data set.

|  |  |
| --- | --- |
| Mean | 278949.67 |
| Standard Deviation | 25583.11 |
| Median | 274988.70 |
| Mode |  |
| Standard Error | 2558.31 |
| Sample Variance | 661106461 |
| Kurtosis | -0.32 |
| Skewness | 0.71 |
| Minimum | 243476.75 |
| Maximum | 347102.32 |
| Range | 103625.57 |
| Sum | 27894967.27 |
| Count | 100 |
| 95% confidence interval (Test Static) | 6719.161 |
| Lower Boundary | 272230.51 |
| Upper Boundary | 285668.83 |

**Out of the top 100 observations that are filtered based on the TotalPay of the employees between 200k-400k, The mean was found to be 278949.67 and the Standard Deviation was found to be 25583.11. The test static for the 95% confidence interval is 6719.161.**

**Lower Boundary: 272230.51**

**Upper Boundary: 285668.83**

**(Rounded the values up to 2 decimal places.)**

**(NOTE: The calculations are done using excel.)**

**Applying lessons learnt from Module 3:**

**1.Conduct the Hypothesis Testing to find weather the Base Pay of the employees exceeds $180000 at the significance level of 0.01.**

Mean(x) = 204861.99

H0: Mean(µ) = 180000

Ha: Mean(µ) > 180000

Confidence level = 99%

Alpha Value = 0.01

n = 100

t-score = (x-µ)/(s/sqrt(n)), t-score = - 4.361 (rounded off to 3 decimal places)

Degrees of freedom (n-1) = 99

Critical t (tc) = -2.364

t < tc

-4.361 < -2.364

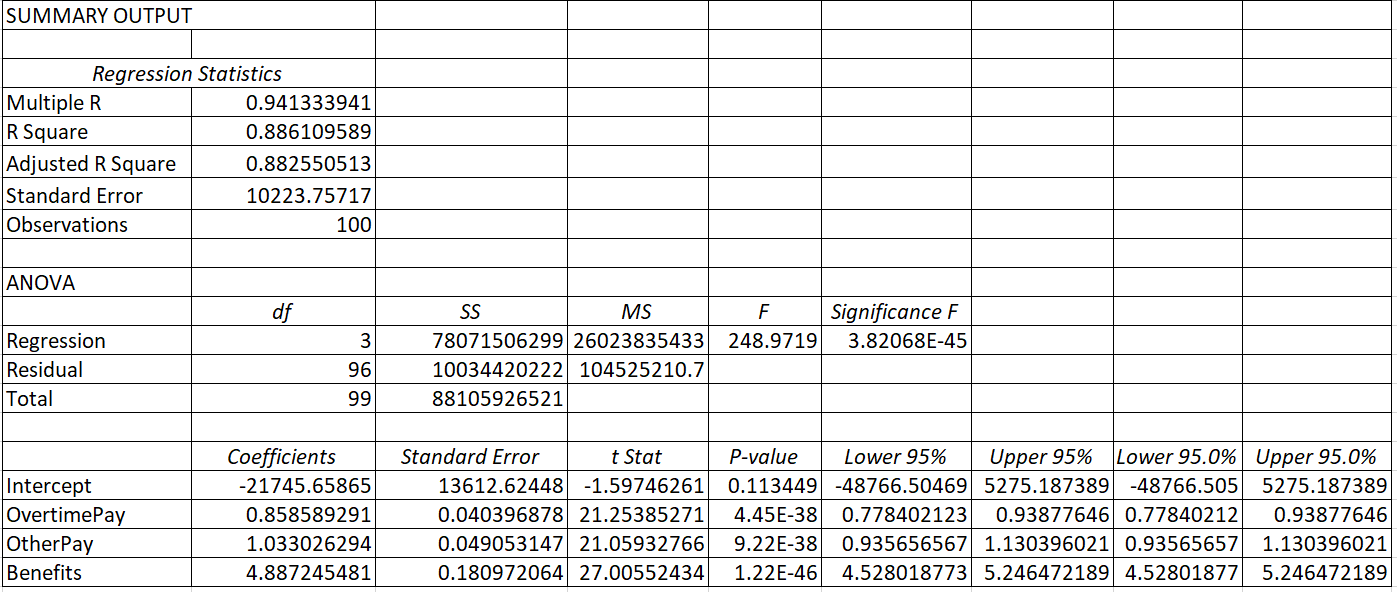
**Since the (t) value is found to be less than the (tc) value, we reject the null Hypothesis. There’s a sufficient evidence to support the claim that the Base Pay of the employees exceeds $180000 at the significance of 0.01.**

**Applying lessons learnt from Module 5:**

**1.Perform the Regression Analysis to predict the TotalPay Benefits and find out the estimated multiple regression equation.**

The Equation of the Multiple Regression Model is given by where β0, β1… βk are the estimated parameters and are the independent variables.

For the dataset taken, The dependent variable is TotalPayBenefits and the independent variables are OvertimePay (x1), OtherPay (x2), Benefits (x3)

****

The equation of the multiple regression equation is given by

**Y = -21745.659 + 0.859 (OvertimePay) + 1.033 (OtherPay) + 4.887 (Benefits)**

A unit increase in the OvertimePay increases the TotalPayBenefits by 0.859, a unit increase in the OtherPay increases the TotalPayBenefits by 1.033, a unit increase in the Benefits increases the TotalPayBenefits by 4.887.

**Adjusted R square for the data is 0.882550513 which indicates we are 88.26% confident that the developed model fits the data well.**

**NOTE:**

* **ALL THE VALUES ARE IN USD ($)**
* **ALL THE CALCULATIONS ARE DONE USING EXCEL.**
* **FOR PROBLEM IN MODULE 1 ( THE ENTIRE DATA SET WAS TAKEN )**
* **FOR REST 3 MODULES THE FILTERED DATA SET IS USED.**