***INSURANCE DATA***

Objective: - Insurance Charges Forecast using Linear Regression

**Types of DATA**

Continuous Data Categorical Data

* In continuous data there are four variable which explain about the data

(Age, Charger, Children, Bmi)

* In categorical data there are two variable which categorise the data

(YES or NO, Gender)

**EDA Analysis: -**

Exploratory Data analysis refers to the critical process of performing initial Investigations on data so as to discover patterns to spot anomalies, to test Hypothesis and to check assumption with the help of summary statistics And graphical representation.

* Descriptive analysis: -

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Age* | *bmi* | *children* | *Charges* |
|  |  |  |  |  |
| Mean | 39 | 31 | 1 | 13270 |
| Standard Error | 0 | 0 | 0 | 331 |
| Median | 39 | 30 | 1 | 9382 |
| Mode | 18 | 32 | 0 | 1640 |
| Standard Deviation | 14 | 6 | 1 | 12110 |
| Sample Variance | 197 | 37 | 1 | 146652372 |
| Kurtosis | -1 | 0 | 0 | 2 |
| Skewness | 0 | 0 | 1 | 2 |
| Range | 46 | 37 | 5 | 62649 |
| Minimum | 18 | 16 | 0 | 1122 |
| Maximum | 64 | 53 | 5 | 63770 |
| Sum | 52459 | 41028 | 1465 | 17755825 |
| Count | 1338 | 1338 | 1338 | 1338 |

* Average age of the person is 39 years who is having 1 children is chargeable rs.13270.
* Minimum age of a person is 18years and maximum age of a person is 64 years.
* Their age which is repeating more 18 years that means company are more focusing 18 years old person.
* **Correlation Matrices:-**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | *age* | *Sex* | *bmi* | *children* | *smoker* | *region* | *Charges* |
| Age | 1 |  |  |  |  |  |  |
| Sex | #DIV/0! | 1 |  |  |  |  |  |
| Bmi | 0.11 | #DIV/0! | 1 |  |  |  |  |
| children | 0.04 | #DIV/0! | 0.01 | 1 |  |  |  |
| smoker | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | 1 |  |  |
| Region | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | 1 |  |
| charges | 0.30 | #DIV/0! | 0.20 | 0.07 | #DIV/0! | #DIV/0! | 1 |

* Here age is correlated to charges that means if age will increase than charges will increase according to that or vice versa. And also Age is not strongly correlated to charges.
* **Graph Visualization: -**

1. **Region Wise smoker**

|  |  |
| --- | --- |
| Region | Smoker |
| northeast | 324 |
| northwest | 325 |
| southeast | 364 |
| southwest | 325 |

**If you see the graph where southeast region people are smoking more so company has two choice either company should charge more or should not focus these area**

* **Men Vs Female Share**

|  |  |
| --- | --- |
| Gender | % Share |
| Female | 49% |
| Male | 51% |

**Here the number of male’s are more as compare to the female (male=51%, female= 49%)**

* **Count of Smoker: -**

|  |  |
| --- | --- |
| Smoker | Count of Smoker |
| No | 1064 |
| Yes | 274 |

**Here smoker is less as compare to Non-Smoker (Smoker=274, Non-Smoker=1064) that is good for the company investment.**

* **Male & Female who don’t smoke**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Gender** | **northeast** | **northwest** | **southeast** | **southwest** | **Grand Total** |
| female | 132 | 135 | 139 | 141 | 547 |
| male | 125 | 132 | 134 | 126 | 517 |

**(If you see the graph where male are smoking more as compare to female’s)**

* **Number of male and female who smoke region wise:-**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Gender** | **northeast** | **northwest** | **southeast** | **southwest** | **Grand Total** |
| female | 29 | 29 | 36 | 21 | 115 |
| male | 38 | 29 | 55 | 37 | 159 |

**No. of male are smoking more in southeast as compare to female.**

**Regression model: -**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance*  *P-level* |  |  |  |
| Regression | 6 | 23693787947 | 3948964658 | 30.49111694 | 1.98624E-34 |  |  |  |
| Residual | 1331 | 1.7238E+11 | 129511971.2 |  |  |  |  |  |
| Total | 1337 | 1.96074E+11 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | -6916.2 | 1758.7 | -3.9 | 0.0 | -10366.4 | -3466.1 | -10366.4 | -3466.1 |
| age | 240.0 | 22.3 | 10.8 | 0.0 | 196.2 | 283.8 | 196.2 | 283.8 |
| Sex | 0.0 | 0.0 | 65535.0 | #NUM! | 0.0 | 0.0 | 0.0 | 0.0 |
| bmi | 332.1 | 51.3 | 6.5 | #NUM! | 231.4 | 432.8 | 231.4 | 432.8 |
| children | 542.9 | 258.4 | 2.1 | 0.0 | 35.9 | 1049.8 | 35.9 | 1049.8 |
| smoker | 0.0 | 0.0 | 65535.0 | #NUM! | 0.0 | 0.0 | 0.0 | 0.0 |
| region | 0.0 | 0.0 | 65535.0 | #NUM! | 0.0 | 0.0 | 0.0 | 0.0 |

*df: -which effect t-test, f-test and adjusted R-Square. It takes into sample size and adjust in normal distribution with lower sample size*

*SS: - sum square*

*MS: -Mean square*

**Analysis: -**

**Here Residual is 1331 that means the predicted charges and actual charges difference is 1331 so company should focus on**

**Python Analysis: -**

* *#Checking the summary of the dataset*

data.describe()

| **age** | **bmi** | **children** | **charges** |
| --- | --- | --- | --- |
| **count** | 1338.000000 | 1338.000000 | 1338.000000 | 1338.000000 |
| **mean** | 39.207025 | 30.663397 | 1.094918 | 13270.422265 |
| **std** | 14.049960 | 6.098187 | 1.205493 | 12110.011237 |
| **min** | 18.000000 | 15.960000 | 0.000000 | 1121.873900 |
| **25%** | 27.000000 | 26.296250 | 0.000000 | 4740.287150 |
| **50%** | 39.000000 | 30.400000 | 1.000000 | 9382.033000 |
| **75%** | 51.000000 | 34.693750 | 2.000000 | 16639.912515 |
| **max** | 64.000000 | 53.130000 | 5.000000 | 63770.428010 |

* **We can see that there is no Null value in our dataset**.

**insurance.isnull().sum()**

age 0

sex 0

bmi 0

children 0

smoker 0

region 0

charges 0

dtype: int64

* **Finding the correlation between charges and other features and arranging them in increasing order**.

region -0.006208

sex 0.057292

children 0.067998

bmi 0.198341

age 0.299008

smoker 0.787251

charges 1.000000

Name: charges, dtype: float64

* **The below plot represents the ditribution of the medical charges which tells us how many patients spend how much money on treatment on average.**

%matplotlib inline

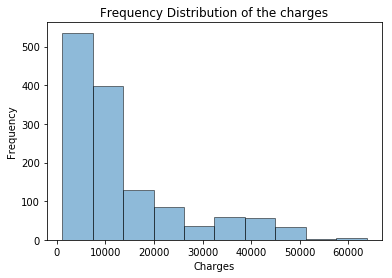
pl.hist(df.charges,bins = 10,alpha=0.5,histtype='bar',ec='black')

pl.title("Frequency Distribution of the charges")

pl.xlabel('Charges')

pl.ylabel('Frequency')

pl.show()

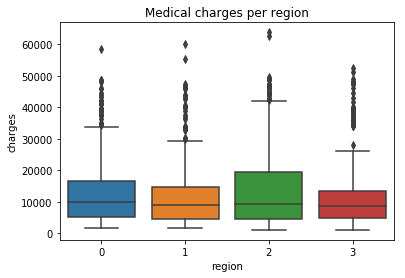


* **The below boxplot shows money spent by people on their treatment in different regions. From the plot we can say that the region doesn't have much impact on medical charges.**

sns.boxplot(x=insurance.region,y=insurance.charges,data=insurance)

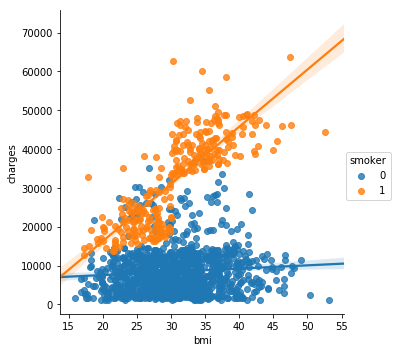
pl.title("Medical charges per region")

pl.show()

****

* **Medical charge increases in case of smoker with the increasing bmi. But in case of non-smokers the increase in bmi doesn't have large impact on the medical charges.**

**sns.lmplot(x="bmi",y='charges',hue='smoker',data=insurance)**

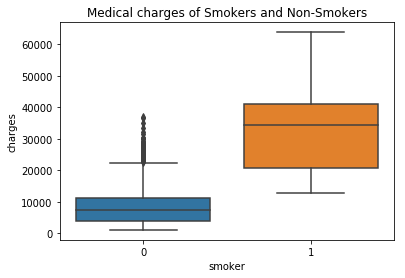
****

* **The boxplot between the medical charges of smokers and non-smokers conveys that those who smoke spend around 4 times more on medicines or treatment as comparerd to those who don't smoke.**

sns.boxplot(x=insurance.smoker,y=insurance.charges,data=insurance)

pl.title("Medical charges of Smokers and Non-Smokers")

pl.show()

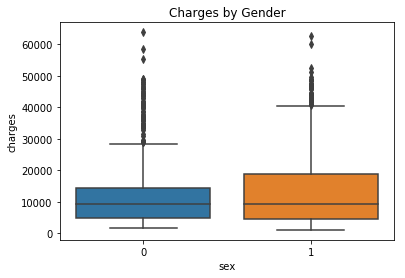


* **The boxplot between sex and charges shows that there is no gender biasing with the medical charges.It doesn't matter whether you are a male or a female the charges remains same for all.**

sns.boxplot(x=insurance.sex,y=insurance.charges,data=insurance)

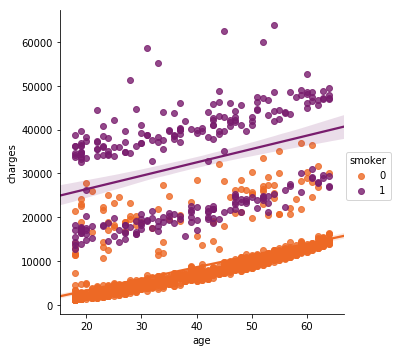
pl.title("Charges by Gender")

pl.show()

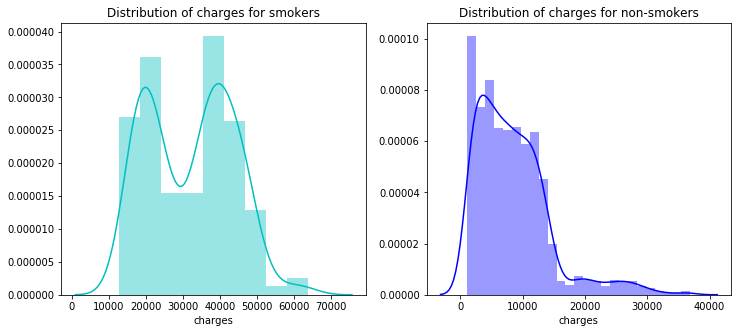


* **In the case of non-smokers,the cost of treatment increases with age, while in case of the smokers,there is not do such dependence. We can also see that age has positive correlation and has larger affect after the smoker that leads to increase in the charges.And it is also a common phenomenon that with increasing age medical expenses willincrease whether you are a smokeror not.**

sns.lmplot(x='age',y='charges',hue='smoker',data=insurance,palette='inferno\_r')



* **These plots represent the distribution of** **medical charges for the smokers** **and non-smokers**.

****