|  |
| --- |
| BUISNESS REPORT ON HEALTH INSURANCE |

|  |
| --- |
| Great learning  Email: info@greatlearning.in  Website: [www.great](http://www.great) learning.com |

|  |  |
| --- | --- |
| R Naveena GLCA  naveeraj111@gmail.com  Email [Email Here] |  |



*Cause and effect of analysis Title on data*

*Of Insurance claims*

1. Perform the Exploratory Data Analysis on the data.

a) Identify the categorical and continuous variable

*Discrete variable*

*A variable that takes only a finite number of real values.*

Categorical variable

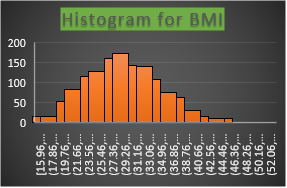
*A variable that codes whether each one in a set of observations is in a particular nominal* category is called a [categorical variable](https://en.wikipedia.org/wiki/Categorical_variable).

*Continuous variable*

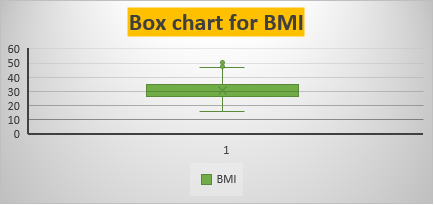
*A continuous variable is a variable whose value is obtained by measuring, i.e., one which can take on*[*uncountable set*](https://en.wikipedia.org/wiki/Uncountable_set)*of values*.

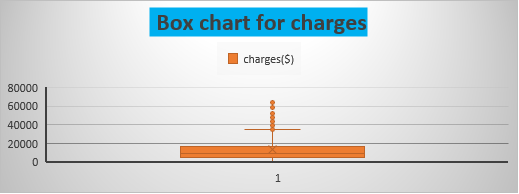
* *Here from using the given data, Sex, smoker, region are categorical variables.*
* *BMI, charges are continuous variable.*
* *Age, children are discrete variables.*

b) Make Histograms and box plots (univariate analysis) for continuous variables and do a correlation analysis (multivariate analysis)

 Graphical user interface, chart

Description automatically generated





|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Correlation (multivariate analysis) | | | | |
|  |  |  |  |  |
|  | *age* | *BMI* | *children* | *Charges ($)* |
| age | 1 |  |  |  |
| BMI | 0.109272 | 1 |  |  |
| children | 0.042469 | 0.012759 | 1 |  |
| charges | 0.299008 | 0.198341 | 0.067998 | 1 |
|  |  |  |  |  |

*1.c) Make relevant Pivot tables and charts for: i. Male/Female ratio and share information on which gender has more smokers*

1. Male/Female ratio and share information on which gender has more smokers

ii. Charges vs Age

iii. Charges vs BMI

iv. Charges for Smokers vs Non-smokers

|  |  |  |
| --- | --- | --- |
| Count of smoker | Smokers |  |
| Sex | no | yes |
| female | 547 | 115 |
| male | 517 | 159 |
|  |  |  |

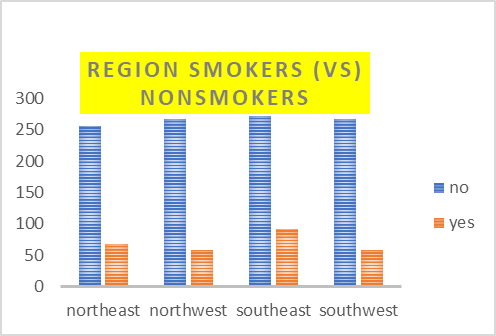
|  |  |  |  |
| --- | --- | --- | --- |
| Ratio of Male/Female | 1.3826087 |  |  |
|  |  |  |  |
|  |  |  |  |
| Here Male gender has more smokers when compared to female gender. | | | |
|  | | |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| * *In this case, make pivot tables for both sex& Smokers.* |  | | | |  |  |
| * *The row of sex contains Male, Female which is used to differentiate them.* | | | | | |  |
| * *The column of smokers contains Yes, No is used to identify the smokers & Non-smokers of the sex.* | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | * *By making this pivot table and recommended chart it shows the,* | | | | | * *Male gender has the more smokers which is 159.* | | | | * *Male/female ratio is 1.3826* |  | |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| C ii). Age& charges: | | | | |  |  |
| |  |  |  | | --- | --- | --- | | * Here, makes the pivot chart for age& charges | |  | | * The age in the row and the charges in the value fields to declare averages. | | | | | * Line chart shows appropriate chart here. |  |  | | * The values are from these charts represent the numerical value. | | | | | | |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C. iii). BMI & charges: | |  | |  | | | |  | | |  | | |  | |  | |
| * In this case, make pivot table for BMI & charges here. | | | | | | | |  | | |  | | |  | |  | |
| * In this pivot table, BMI in row & averages of the charges in the value fields. | | | | | | | | | | | | | |  | |  | |
| * Use Area chart here because huge data from them. So, area chart seems to be appropriate here... | | | | | | | | | | | | | | | | | |
| * These two are also indicates the numerical values... | | | | | | | |  | | |  | | |  | |  | |
|  |  | |  | | | |  | | |  | | |
|  | | | | | | |  | | |  | | |
|  | | | | | | | | | | | | |
|  | | |  | | | |  | | |  | | |
|  | | | | | | | | | |  | | |
| C. iv). Charges of smoker’s vs non-smokers: | | | | | |  | | |  | | |  | | |  | |
| * *Use pivot table for charges & smokers.* | | | | |  |  | | |  | | |  | | |  | |
| * *Smokers in the row use to differentiate the smokers & Non-smokers.* | | | | | | | | | | | |  | | |  | |
| * *Charges in the value field shows the averages...* | | | | | |  | | |  | | |  | | |  | |
| * *Here, use Bar chart to differentiate the ranges of the charges of smokers vs Non-smokers.* | | | | | | | | | | | | | | | | |

1.d) Region-wise smokers vs Non-smokers analysis with one or more pivot table and charts

|  |  |  |
| --- | --- | --- |
| Count of smoker | Column Labels |  |
| region | No | yes |
| northeast | 257 | 67 |
| northwest | 267 | 58 |
| southeast | 273 | 91 |
| southwest | 267 | 58 |



1.e) Region-wise charges for smoker’s vs non-smokers:

|  |  |  |
| --- | --- | --- |
| Average of charges($) | Column Labels |  |
| Row Labels | no | yes |
| northeast | 9166 | 29674 |
| northwest | 8556 | 30192 |
| southeast | 8032 | 34845 |
| southwest | 8019 | 32269 |

This shows that southeast region has more smokers.

1.f) Has charges got something to do with the number of dependents

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Correlation of number of dependents & charges | | | | | |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  | *children* | *charges ($)* |  |  |  |
|  |  | children | 1 |  |  |  |  |
|  |  | charges($) | 0.067998 | 1 |  |  |  |
|  |  |  |  |  |  |  |  |
|  | We can say it is an independent. Because it is positive corelation | | | | | | |

1. g) Do a similar dependants-charges analysis, Region-wise

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Average of charges ($) | Column Labels | |  |  |  |  |
| Row Labels | 0 | 1 | 2 | 3 | 4 | 5 |
| northeast | 11626 | 16310 | 13615 | 14410 | 14485 | 6979 |
| northwest | 11324 | 10230 | 13464 | 17786 | 11347 | 8966 |
| southeast | 14310 | 13687 | 15728 | 18450 | 14451 | 10115 |
| southwest | 11939 | 10406 | 17483 | 10402 | 14933 | 8444 |
|  |  |  |  |  |  |  |

1. h) Do at least one more pivot table and chart of your own choice on the remaining variables

|  |  |  |
| --- | --- | --- |
| Average of charges ($) | Column Labels |  |
| Row Labels | no | yes |
| female | 8762.2973 | 30678.99628 |
| male | 8087.204731 | 33042.00598 |



1. i) Give your understanding from the patterns observed in point (b)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Univariate analysis: | | | |  |  |  | |  | | | |  |  |  |
|  |  | |  |
| * *Univariate analysis is the simplest form of statistical analysis...* | | | | | | |
| * *It can be Inferential or descriptive. (One variable only involved).* | | | | | | |
|  |  | |  |
| Continuous variables: | | | |  |  |  | |  | | | |  |  |  |
|  |  | | | | | | |  | | | |  |  |  |
| * *If a variable over a non-empty range of the real numbers is continuous,* | | | | | | | | | | |
| * *then it can take on any value inn that range.* | | | | | | | | |  |  |
| * *So, we declare that BMI & Charges are continuous variable here.* | | | | | | | |  | | | |  |  |  |
| * *Here, we should make a histogram and box charts for both these variables.* | | | | | | | | | | | |  |  |  |
|  |  |  | |  |  |  | |  | | | |  |  |  |
| Multivariate analysis: | | | |  |  |  | |  | | | |  |  |  |
|  | *Is the analysis of more than one statistical outcome variable at a time which is more appropriate*  *.* | | | | | | | | | | | | | |
| Correlation Coefficient: | | | |  |  |  | |  | | | |  |  |  |

* Here, Variables moves in one relation to another. They have (positive correlation, inverse correlation).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| * *Here, Age, BMI charges, children are having numerical values.* |  |  |  |  |
| * *So, we make the correlation analysis for all these numerical values.* | |  |  |  |

1.j) Give your interpretation for observations made in point (c)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 1.c) i). Male /Female ratio: | |  |  |  |  |  |
| * *In this case, make pivot tables for both sex & smokers* | | | |  |  |  |
| * *The row of sex contains Male, Female which is used to differentiate them.* | | | | | |  |
| * *The column of smokers contains Yes, No is used to identify the smokers & Non smokers* | | | | | | |
| * *of the sex.* |  |  |  |  |  |  |
| * *By making this pivot table and recommended chart it shows the,* | | | | |  |  |
| * *Male gender has the more smokers which is 159.* | | | |  |  |  |
| * *Male/female ratio is 1.3826* | |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C ii). Age& charges: | | |  |  |  |  |  |  |
| * *Here, makes the pivot chart for age& charges* | | | | |  |  |  |  |
| * *The age in the row and the charges in the value fields to declare averages.* | | | | | | |  |  |
| * *Line chart shows appropriate chart here...* | | | |  |  |  |  |  |
| * *The values are from these charts represent the numerical value ...* | | | | | |  |  |  |
|  |  |  |  |  |  |  |  |  |
| C. iii). BMI & charges: | | |  |  |  |  |  |  |
| * *In this case, make pivot table for BMI & charges here...* | | | | |  |  |  |  |
| * *In this pivot table, BMI in row & averages of the charges in the value fields...* | | | | | | |  |  |
| * *Use Area chart here because huge data from them...So area chart seems to be appropriate here...* | | | | | | | | |
| * *These two are also indicates the numerical values...* | | | | |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| C. i v). Charges of smoker’s vs non-smokers: | | | | |  |  |  |  |
| * *Use pivot table for charges & smokers.* | | | |  |  |  |  |  |
| * *Smokers in the row use to differentiate the smokers & Non-smokers.* | | | | | | |  |  |
| * *Charges in the value field shows the averages...* | | | | |  |  |  |  |
| * *Here, use Bar chart to differentiate the ranges of the charges of smoker’s vs Non-smokers.* | | | | | | | | |

2. Edit the data as following, to obtain dummy variables:

a) Sex: Replace all the “Males” with “1” and “Females” with “0”, creating numerical entries for gender this way will help you do analysis further. You can use the “Replace with Match entire cell content” option. Do a replace all to save time.

*From the given data replace all the “Male” in to”1”,”Female” in “0” using the IF condition as =IF(A5="male", 1,0). The entire gender will changes.*

2. b) Smoker: Replace all the “Smokers” with “1” and “Non-smokers” with “0”.

*From the given data replace all the “Smokers” in to”1”,”Non smokers” in “0” using the IF condition as =IF(A5="Smokers", 1,0). The entire gender will changes.*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |

c) Region: We always create one less category column for the dummy data w.r.t the categories available for that original variable. So for Region, we will create three dummy columns, assuming “Northeast” as zero and omit the column for it. Now create three columns for “northwest”, “Southeast”, “Southwest”. Whichever row has “northwest” region as an entry will take “1” as an entry otherwise “0” in “northwest” column. Similarly, in the “Southeast” column, whichever row had “southeast” as an entry will take “1” as the new entry and “0” for the rest of the column (Southeast). Do a similar operation on the “Southwest” column. Please refer to the below image for your understanding,

*From the given data replace all the “northwest” in to”1 using if condition as* *=IF(A5="northwest",1,0)* *and replace all “Southeast” in to 1 using IF condition as=IF(A5="southeast",1,0) also replace all “Southwest”* *into 1 by using IF condition as=IF(A5="southwest",1,0)..So the entire columns of the region are change.*

1. Do a descriptive summary analysis for the edited data. Perform a Multiple Linear Regression analysis to identify which variables decide the insurance charges/billed insurance claim. Give your interpretation for the above analysis, do another set of regression analysis by dropping insignificant Variables, if needed.
   * *From the given data source, we want to change all the character into numerical*

* *values by using IF condition. After that we want to find the Descriptive statistics by*
* *using Data Analysis source.so, that we can enable the functions of the descriptive process.*
* *We will obtain the table* *with the same if condition data. Using the data analysis, regression process has been*
* *Made with the x range as age, sex, BMI, children ,smokers ,southwest ,northwest ,southeast with charges as Y range coefficient values, significant etc. with Absolute value.*
* *In this regression process the x range will be Age, BMI ,children ,smokers ,southwest, southeast and the charges(y),by using data analysis and regression can be obtained.*

|  |  |
| --- | --- |
|  | Descriptive Statistics for all numerical variables: |



|  |  |
| --- | --- |
|  | *Coefficients* |
| Intercept | -11938.53858 |
| age | 256.8563525 |
| Sex | -131.3143594 |
| bmi | 339.1934536 |
| children | 475.5005451 |
| Smoker | 23848.53454 |
| Northwest | -352.9638994 |
| Southeast | -1035.022049 |
| Southwest | -960.0509913 |

Multi linear Regression:

|  |  |
| --- | --- |
| *Regression Statistics* | |
| Multiple R | 0.866552384 |
| R Square | 0.750913035 |
| Adjusted R Square | 0.74941364 |
| Standard Error | 6062.102289 |
| Observations | 1338 |
|  |  |

Significant Regression:

|  |  |
| --- | --- |
|  | *Coefficients* |
| Intercept | -12165.38244 |
| age | 257.0063906 |
| bmi | 338.6413347 |
| children | 471.5441444 |
| Smoker | 23843.87493 |
| Southeast | -858.4696418 |
| Southwest | -782.7452298 |

|  |  |
| --- | --- |
| *Regression Statistics* | |
| Multiple R | 0.866476426 |
| R Square | 0.750781397 |
| Adjusted R Square | 0.749657948 |
| Standard Error | 6059.146461 |
| Observations | 1338 |

These are the regression for the significant variables.