|  |
| --- |
| BUSINES  REPORT ON HEALTH INSURANCE |

|  |
| --- |
| Great Learning  Email: [info@greatlearnings.in]  Website: [www.mygreatlearnings.com] |

|  |  |
| --- | --- |
| Swetha, GLCA-DA  Email [swetha.laksmi99@gmail.com] |  |



Contents

Question no.1 (a) to (j)2 -11

Qn no.1 (c (i) to (iv)………………………………………………………………………………………….. 5-7

Question no.212

Question no.3…………………………………………………………………………………………………………..1

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 continuos variables?

|  |
| --- |
| TYPES OF DATA |
| *1)Categorical* |
| *2)Numerical* |
| *a)Discrete* |
| *b)Continuous* |

CATEGORICAL: Categorical variables represent types of data which may be divided into groups.[Here in the data “SEX,SMOKER &REGION” are categorical].

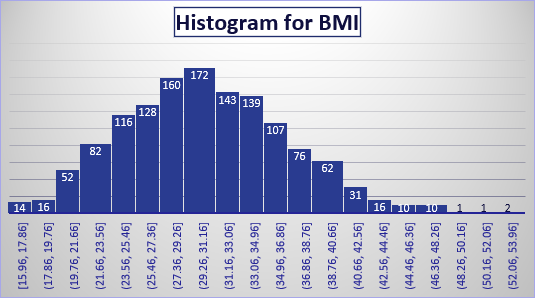
DISCRETE: Discrete data is a numerical type of data that includes whole, concrete numbers with specific and fixed data values determined by counting.

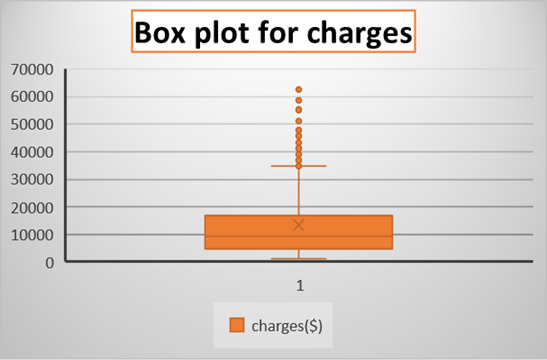
[Here in the data “AGE & CHILDREN “are discrete]

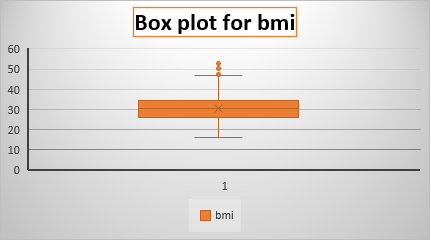
CONTINUOUS:  Continuous data includes complex numbers and varying data values measured over a particular time interval.[Here in the data ,”BMI &CHARGES” are continuous]

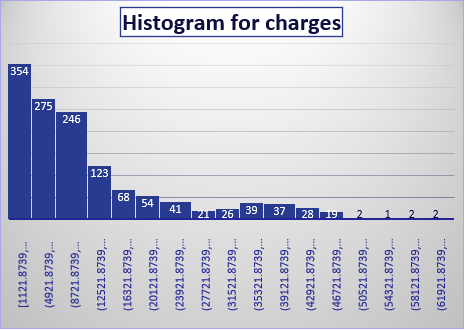
b)Histograms and box plots for (univariate analysis ) for continuous variables and do a correlation analysis (multivaraiate analysis)

Univariate analysis is the simplest form of analysing data.we are considering the BMI,CHARGES($) as continuous to make histograms and box plots.









Multivariate analysis method is used to determine the effects of qualitative variables.correlation is used to find the relation between properties related to age,bmi,children,charges.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | age | bmi | children | charges($) |
| age | 1 |  |  |  |
| bmi | 0.109272 | 1 |  |  |
| children | 0.042469 | 0.012759 | 1 |  |
| charges($) | 0.299008 | 0.198341 | 0.067998 | 1 |

c)pivot tables of relevant data

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

|  |  |  |
| --- | --- | --- |
| Count of smoker | Smoker |  |
| Sex | no | yes |
| female | 547 | 115 |
| male | 517 | 159 |
|  | MALE/FEMALE RATIO | 1.3826 |

Make a pivot for sex and smoker. By dragging the ' sex' field in ROW area,'smoker' field in COLUMN area,'smoker'field in VALUE AREA can get the COUNT OF SMOKER.

By finding the Male to female ratio,we can come to the conclusion that MALE has higher ratio of smoking[(i.e) ratio is 1.38]

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

ii. Charges vs Age

Make a pivot for age and charges.By dragging the ' age' field in ROW area,'charges' field in VALUE AREA can get the AVERAGE OF CHARGES.This line chart represent the numerical data.

ii. Charges vs BMI

Make a pivot for Bmi and charges.By dragging the ' BMI' field in ROW area, 'charges' field in VALUE AREA can get the AVERAGE OF CHARGES.This Area chart represent the numerical data.

v. Charges for Smokers vs Non-smokers

|  |  |
| --- | --- |
| Smoker | Average of charges($) |
| No | 8434.268298 |
| Yes | 32050.23183 |

Make a pivot for smoke and charges.By dragging the ' Smoke' field in ROW area, 'charges' field in VALUE AREA can get the AVERAGE OF CHARGES.This Bar chart represent the numerical data.

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 | |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

e) Region-wise charges for smokers vs non-smokers

|  |  |  |
| --- | --- | --- |
| Count of smoker | Smoker |  |
| Region | no | yes |
| northeast | 257 | 67 |
| northwest | 267 | 58 |
| southeast | 273 | 91 |
| southwest | 267 | 58 |

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

|  |  |  |
| --- | --- | --- |
|  | *children* | *charges($)* |
| children | 1 |  |
| charges($) | 0.067998 | 1 |

Since it has positive correlation,we will say that it is a dependent

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Average of charges($) | Children |  |  |  |  |  |
| Region | 0 | 1 | 2 | 3 | 4 | 5 |
| northeast | 11626.46266 | 16310.2064 | 13615.15272 | 14409.9133 | 14485.19312 | 6978.9735 |
| northwest | 11324.37092 | 10230.25631 | 13464.31469 | 17786.16067 | 11347.01873 | 8965.7958 |
| southeast | 14309.86838 | 13687.04197 | 15728.47062 | 18449.84602 | 14451.02397 | 10115.442 |
| southwest | 11938.50499 | 10406.48495 | 17483.48556 | 10402.44226 | 14933.26053 | 8444.1586 |

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

|  |  |  |
| --- | --- | --- |
| Sum of charges($) | Sex |  |
| Smoker | female | male |
| no | 4792976.623 | 4181084.846 |
| yes | 3528084.572 | 5253678.95 |
|  |  |  |

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

Q&A-1(i)OBSERVATION FROM THE POINT IN (b):

UNIVARIATE ANALYSIS:

Univariate analysis is the simplest form of analyzing data. Uni means one, so in other words the data has only one variableUnivariate data requires to analyze each variable separately.

CONTINUOUS VARIABLE:

A continuous variable is defined as a variable which can take an uncountable set of values or infinite set of values.So,We consider the Bmi &Charges as continuous variables.Also we find the histogram and box plot for each variables.

MULTIVARIATE ANALYSIS:

Multivariate analysis is used to study more complex sets of data than what univariate analysis methods can handle.Correlation coefficients are used to measure how strong a relationship is between two variables.Age,Bmi,children,Charges is an numerical value,so that we can make the correlation between them.

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

OBSERVATION FROM THE POINT IN (C)

(i)Male to female ratio

Make a pivot for sex and smoker.By dragging the ' sex' field in ROW area,'smoker' field in COLUMN area,'smoker'field in VALUE AREA can get the COUNT OF SMOKER.By finding the Male to female ratio,we can come to the conclusion that MALE has higher ratio of smoking[(i.e) ratio is 1.38].This Bar chart , represent the visual data of male and female ratio easily.

(ii)Age vs Charges($)

Make a pivot for age and charges.By dragging the ' age' field in ROW area,'charges' field in VALUE AREA can get the AVERAGE OF CHARGES.This line chart represent the numerical data.

(iii)BMI vs Charges($)

Make a pivot for Bmi and charges.By dragging the ' BMI' field in ROW area, 'charges' field in VALUE AREA can get the AVERAGE OF CHARGES.This Area chart represent the numerical data.

(iv) Smoker vs vs charges($)

Make a pivot for smoke and charges.By dragging the ' Smoke' field in ROW area, 'charges' field in VALUE AREA can get the AVERAGE OF CHARGES.This Bar chart represent the numerical data.

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

From the given data replace all the “Males “into “1” and “Females”with ‘0’.

Using the formula of IF condition as =if(A2=”Male”,1,0).The entire gender will change.

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

From the given data replace all the “smokers “into “1” and “Non smokers”with ‘0’.Using the formula of IF condition as =if(A2=”Smokers”,1,0).The entire smoker will change.

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 of category column.Southwest,NorthWest,SouthEast are, replaced with”1”in individual manner.Other regions are represented by”0”.

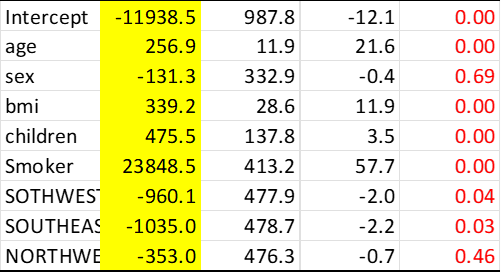
3. Do a descriptive summary analysis for the edited data. Perform a Multiple Linear Regression analysis to identify which variables decide the insurance charges/billed insurance claim. Give your interpretation for the above analysis, do another set of regression analysis by dropping insignificant variables, if needed.

i) 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.



ii)By Descriptive Summary Output Regression for insurancecharges($).

With the same if condition data.Using the data analysis,regression process have been made with the x range age,sex,bmi,chikdren,smokers,southwest,northwest,southeast with charges as Y range coeeficient,P values,significant etc with Absolute value.

iii) In this regression process the x range will be age,bmi,chikdren,smokers,southwest,southeast and the charges(y),by using data analysis and regression can be obtained.