# HEALTH INSURANCE

BUSINESS REPORT

K M Jayasuryan
GL-MAR-23-37 |

**Index:**



**Executive summary:**

This report delves into a dataset encompassing categorical and continuous variables to provide key insights for insurers.

Age emerges as a significant factor positively linked to medical charges, with smoking status notably influencing costs.

Linear regression identifies a model with age, BMI, children, and smoking status as the most effective charge predictor.

Insurers are advised to leverage demographic insights for refined pricing, tailoring premiums based on age and smoking status.

This approach, supported by the recommended linear regression model, enhances data-driven decisions for fair and competitive insurance pricing.

The report underscores age, smoking status, and BMI as pivotal considerations, offering actionable insights for strategic decision-making in the insurance industry.

## Introduction:

In this health insurance claim analysis, the objective is to uncover influential health parameters affecting insurance charges. With this dataset of customer claims and the analysis involves conducting exploratory data analysis through use of statistical tools.

The objective is to discern the key variables that significantly influence insurance charges, contributing to a comprehensive understanding crucial for making informed business decisions in the insurance domain

## Data dictionary:

| Attribute | Description |
|---|---|
| Age | Age of the customer/claimant who has claimed insurance for medical treatment charges |
| Sex | Gender of the customer/claimant |
| BMI | Health parameter: person's weight in kilograms divided by the square of height in meters |
| Children | No. of children the claimant has |
| Smoker | Whether the claimant smokes or not |
| Region | Region to which the claimant belongs |
| Charges | The exact medical charges for which the claimant has claimed insurance |

## Identification of categorial and continuous variables:

Categorial variables are:

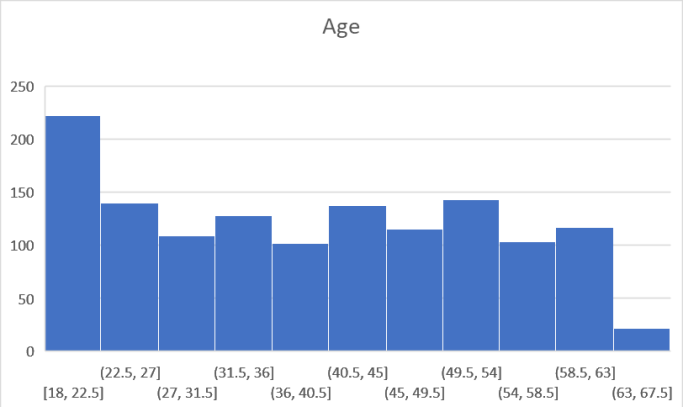
1. Sex: Representing gender as male or female.
2. Smoker: Indicating whether individual is smoker or not.
3. Region: Variable category representing different regions.

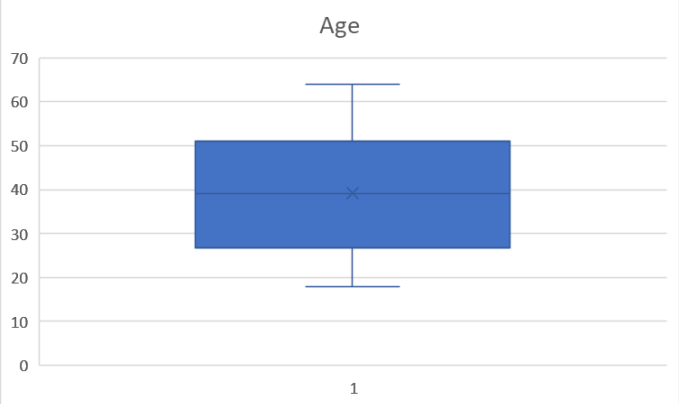
Continuous variable:

1. Age: Representing age of individuals.
2. BMI: Stands for body mass index which is a continuous variable.
3. Children: Represents number of children covered by insurance.
4. Charges ($): Represents insurance charges.

## Explanation of Histogram and Box Plot for continuous variable:

Age: Histogram & Box Plot:




Inference of age:(histogram)

1) In this distribution, we can observe that it is approximately normally distributed. The peakedness is not very high, indicating a platykurtic distribution.

2)There is a higher concentration of individuals within the age range of 18 to 22.5 in the dataset.

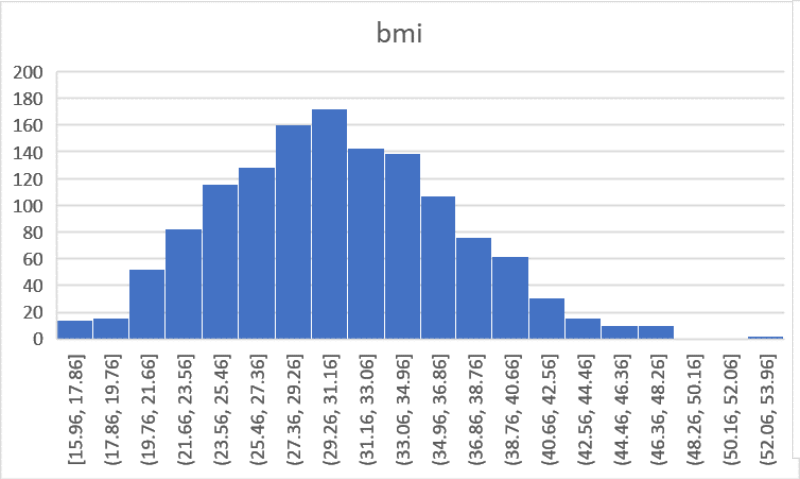
Box plot:

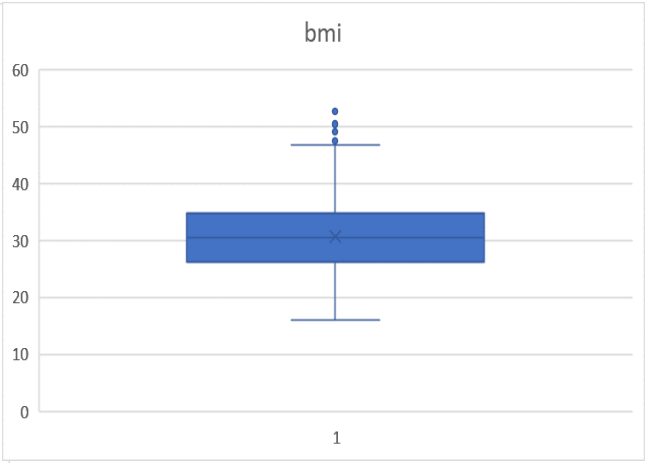
1)In this box plot we can observe that median(q2) is 39.207. Which explains that the middle age for people making insurance claim is around 39 years.

2) IQR which is the spread between Q1 and Q3 is 24.25.

3) The whiskers represent the spread of the data beyond the interquartile range (IQR), indicating the minimum and maximum values. In this case, the minimum is 18, and the maximum is 64.

BMI: Histogram & Box Plot:




Inference of BMI:(histogram)

1)In this distribution, we can observe little peaked and distribution is near to normal for each variable.

2) There are 1% number of people who are under weight.

3) Nearly 50% of people are having BMI status as overweight and obesity.

4) Remaining people nearly 47% are having normal weight.

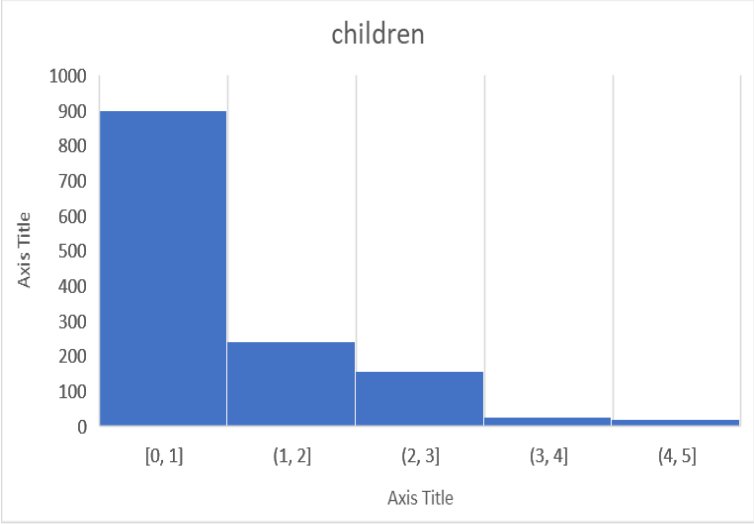
Box plot:

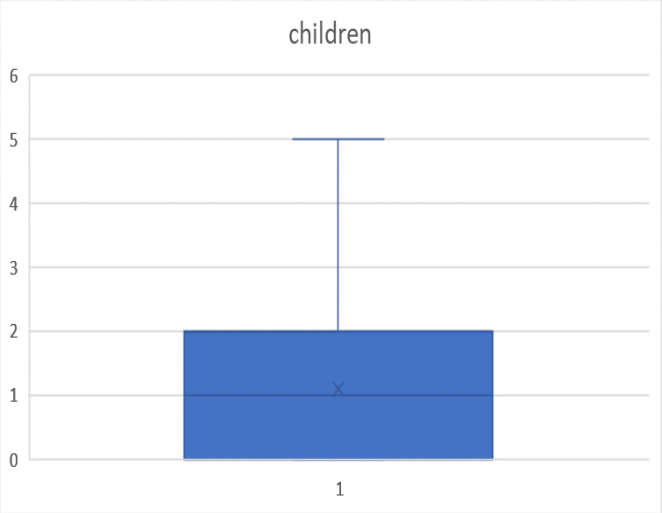
1)In this box plot we can observe that median(q2) is 30.6633. Which means half of the individuals claiming insurance benefits have a BMI below 30.663 and other half have a BMI above this value.

2) IQR which is the spread between Q3 and Q1 is 8.4275.

3) The whiskers represent the spread of the data beyond the interquartile range (IQR), indicating the minimum and maximum values. In this case, the minimum is 15.96, and the maximum is 47.41 are the BMI status of people.

**Children** (Number of dependents): Histogram and Box Plots:




Inference: Histogram

1) We can observe that distribution of frequency is more on one side of distribution and its heavily peaked and trialed to the right which means positively skewed.

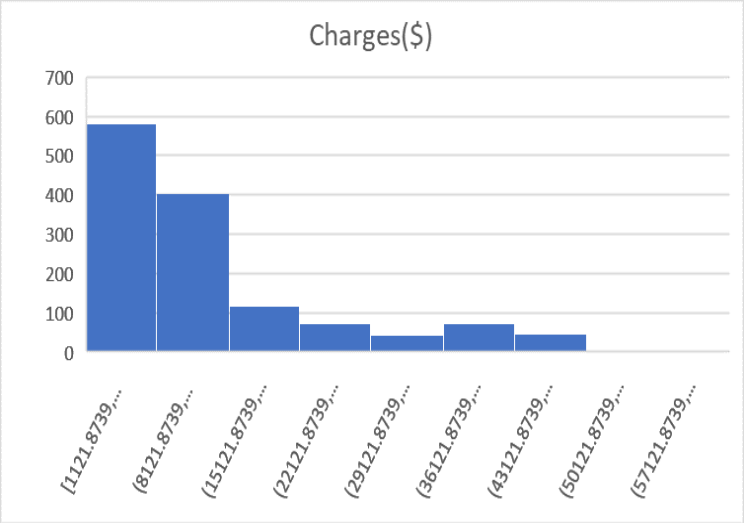
2) Here we can say that majority of insured people are having 0-1 dependent and where a smaller number of people are having 4-5 dependent.

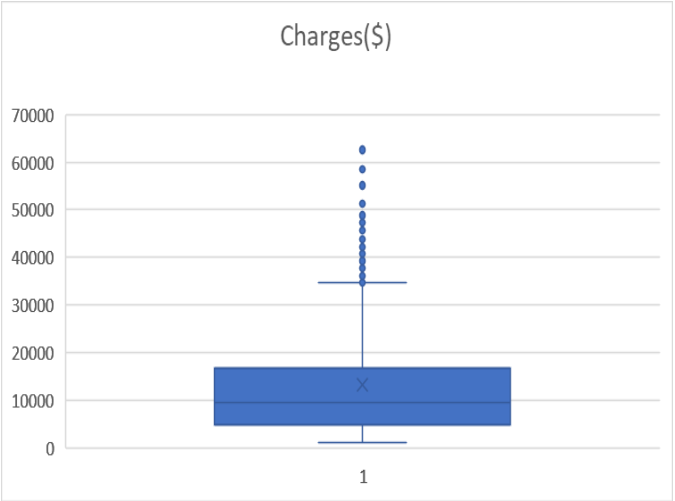
Box Plot:

1) People who claimed insurance benefits are evenly split between those with more than one dependent and those with 1 or fewer dependents.

2) Very less people claimed insurance having nearly 5 dependents.

**Insurance Charges:** Histogram & Box Plot:




Inferences: Histogram:

1) We can observe that distribution of frequency is more peaked which means leptokurtic and it is trialed to the right which is positively skewed.

2) Where insurance company pay high insurance charges are less in number.

3) And insurance company pay less than $155000 are approximately more in number.

Box Plot:

1) Half of the people who claimed insurance benefits by insurance company are charged greater than or less than $ 9382.033.

2) Few people have received insurance charges less than $1121.87 & more than $346792.16.

**Correlation of Variables**

| | *age* | *bmi* | *children* | *charges($)* |
|---|---|---|---|---|
| age | 1 | | | |
| bmi | 0.109272 | 1 | | |
| children | 0.042469 | 0.0127589 | 1 | |
| charges($ | 0.299008 | 0.19834097 | 0.067998227 | 1 |

Here we can observe that age, BMI, children are positively correlated with charges.

Age and charges show a relative correlation than other variables.

**Explanation of Relevant pivot table and charts:**

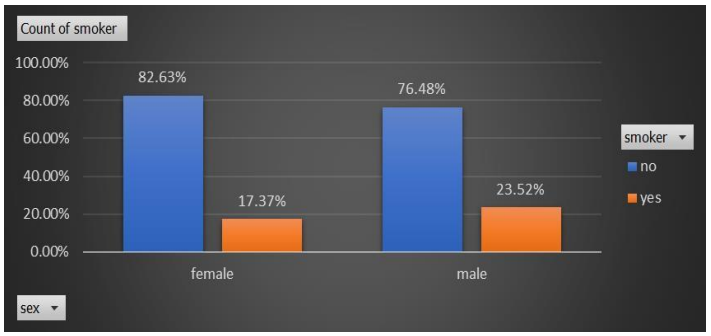
- Male/Female ratio and share information on which gender has more smokers.

1) In this population, 17.37% of females and 23.52% of males are smokers.

2) For example, in a group of 100 people, 20 individuals would be smokers, while 80 would be non- smokers.

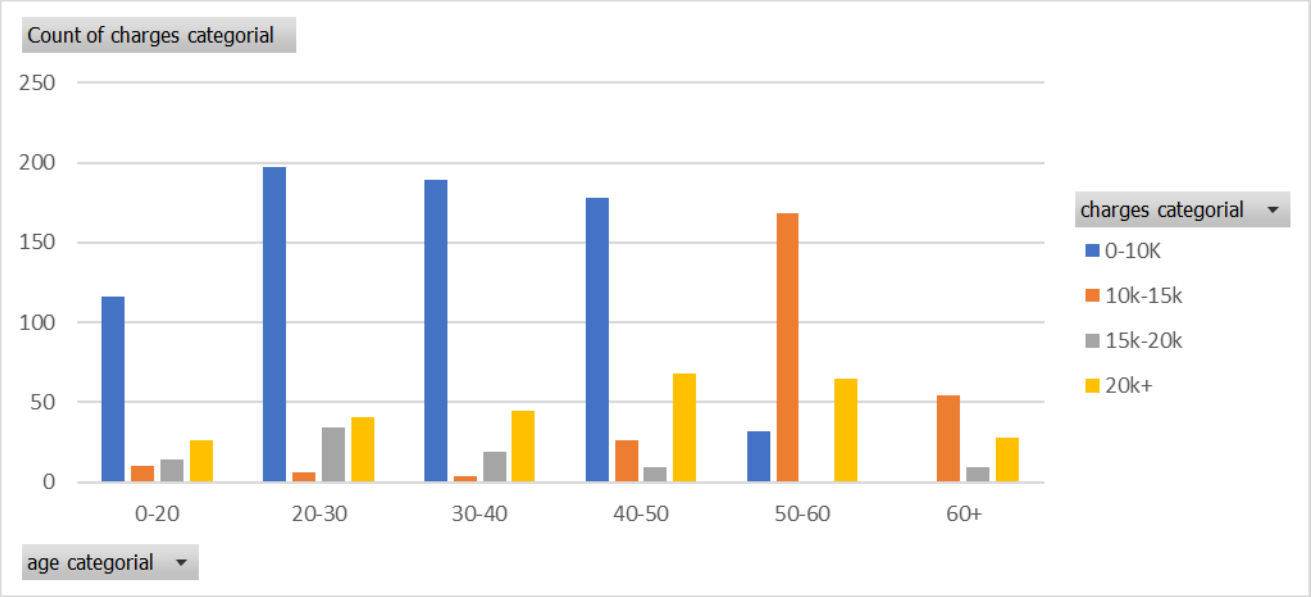
3) According to the data there are more male smokers than female smokers.

| Count of smoker | Column Labels | | |
|---|---|---|---|
| Row Labels | no | yes | Grand Total |
| female | 82.63% | 17.37% | 100.00% |
| male | 76.48% | 23.52% | 100.00% |
| **Grand Total** | **79.52%** | **20.48%** | **100.00%** |




- Charges vs Age:

| Count of charges categ | Column Labels | | | | |
|---|---|---|---|---|---|
| Row Labels | 0-10K | 10k-15k | 15k-20k | 20k+ | Grand Total |
| 0-20 | 116 | 10 | 14 | 26 | 166 |
| 20-30 | 197 | 6 | 34 | 41 | 278 |
| 30-40 | 189 | 4 | 19 | 45 | 257 |
| 40-50 | 178 | 26 | 9 | 68 | 281 |
| 50-60 | 32 | 168 | | 65 | 265 |
| 60+ | | 54 | 9 | 28 | 91 |
| **Grand Total** | **712** | **268** | **85** | **273** | **1338** |

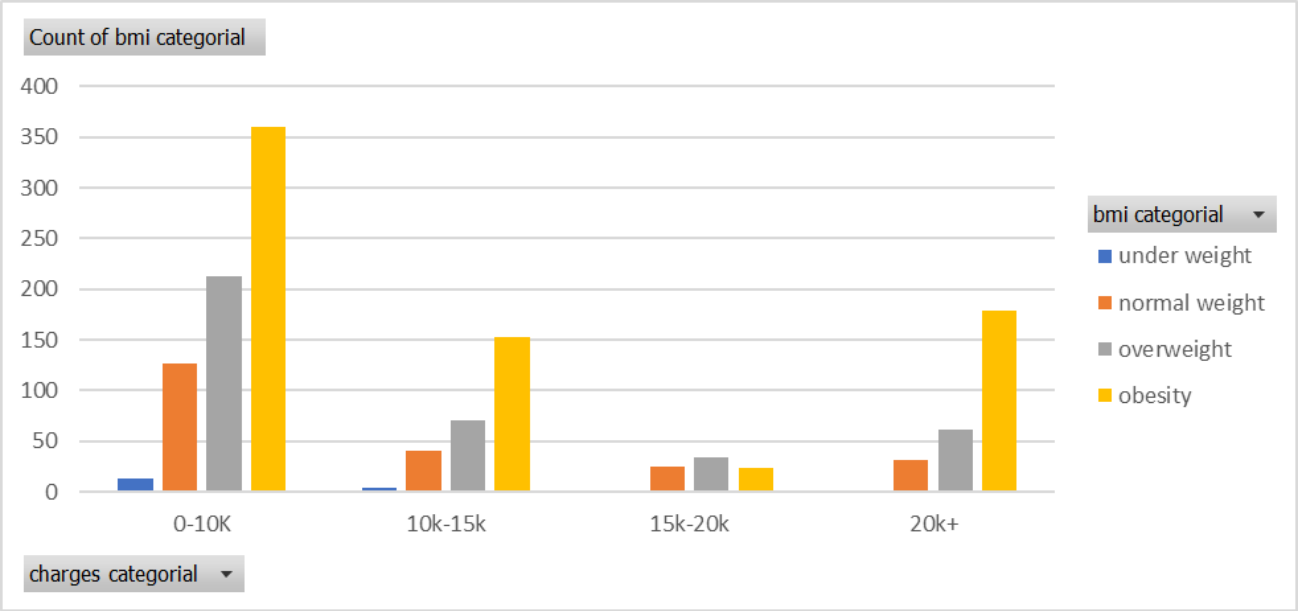



a) In the above data most of the people between the age of 0 and 20 insurance charges are paid by the insurance company in the range of 0-10k like around 116 peoples.
b) Remaining 50 people charged in the range of 10k-20k+
c) People between the age category of 20 and 30 also claimed insurance charges of 0-10k more comparing to other insurance charge category.
d) Out of 278 people 197 people received insurance charges in the category of 0-10k and remaining 81 people claimed in the category of 10k-20k+.
e) Same goes for the people aged category of 30-50 received the insurance category of 0-10k more than other insurance category.
f) People aged 50-60 out of 265 people 168 people received the insurance claim of the category 10k-15k and not a single person claimed in the category of 15k-20k and around 65 people claimed the insurance category of 20k+.
g) People belonging to the age category of 60+ there are 91 people claimed the insurance where 54 people claimed around 10k-15k of insurance category and no one claimed in the category of 0-10k.

With these key findings we can say that in the insurance category of 0-10k most of the individual medical costs have been billed by health insurance company.
Around 53.2% of bills have been claimed in the category of 0-10k and remaining 48.8% of bills have claimed by remaining category from 10k-20k+.

- **Charges vs BMI:**

| Count of bmi categorial | Column Labels | | | | |
|---|---|---|---|---|---|
| Row Labels | under weight | normal weight | overweight | obesity | Grand Total |
| 0-10K | 13 | 126 | 213 | 360 | 712 |
| 10k-15k | 4 | 40 | 71 | 153 | 268 |
| 15k-20k | 2 | 25 | 34 | 24 | 85 |
| 20k+ | 1 | 31 | 62 | 179 | 273 |
| **Grand Total** | **20** | **222** | **380** | **716** | **1338** |




a) In this dataset we can observe that among 1338 people 53.51% of people are in the category of obesity (having excess body mass index),28.4% of people are in the category of overweight, 16.5% of people are in the category of normal weight and 1.49% of people are in the category of underweight.

b) Talking about under weight category people out of 20 people 13 have claimed their hospital bill in the insurance category of 0-10k and remaining 7 people have claimed their bill in other categories of insurance.

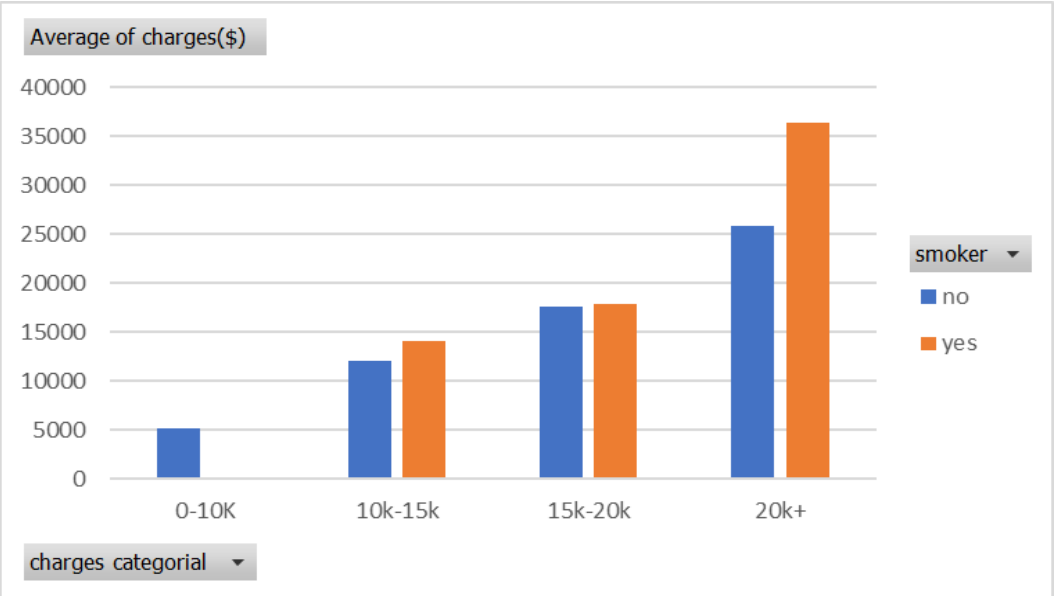
c) People belonging to normal weight category there are around 222 people and maximum of individual medical cost billed by health insurance is in the insurance category of 0-10k (i.e,56.7% of people in normal weight category) and remaining 45.3% of people claimed in remaining category of insurance.

d) Even in overweight category around 56.05% of individual medical cost billed in the category of 0-10k.
e) People in the obesity category total 716 individuals, with 50.2% of them having billed their medical expenses in the 0-10k range. Approximately 21.3% have billed in the 10k-15k range, and 25% of individuals have billed in the 20k+ category.

By the findings we can say that individual falling in the category of obesity have billed their medical cost by health insurance in more number compare to other BMI category individuals.

- **Charges for smoker's vs non-smokers:**

| Average of charges($) | Column Labels | | |
|---|---|---|---|
| Row Labels | no | yes | Grand Total |
| 0-10K | 5207.231751 | | 5207.231751 |
| 10k-15k | 12122.49677 | 14063.51021 | 12173.19488 |
| 15k-20k | 17550.0729 | 17844.04777 | 17740.29193 |
| 20k+ | 25836.69868 | 36329.69964 | 33985.10968 |
| **Grand Total** | **8434.268298** | **32050.23183** | **13270.42227** |



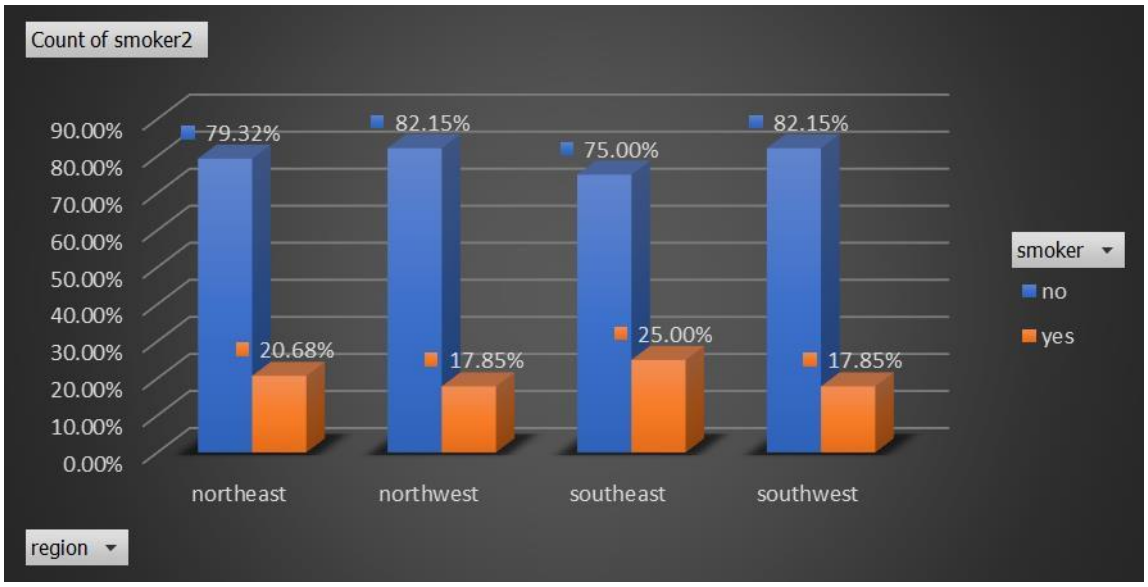

a) In this dataset we observe that individual who smoke have been billed around average of $32050.23. Whereas non-smokers have billed on average of $8434.26.

b) Where non-smokers around 79.52% of individuals have claimed their medical bill in the average charge of $8434.26 and remaining 20.48 individuals who are smokers have claimed their medical bill in the average charge of $ 32050.231.

Based on this data, it appears that smokers, on average, have claimed higher medical bills, but with a smaller number of individuals. Conversely, non-smokers, on average, have lower medical bills, but with a higher number of individuals making claims.

## Analysis based on region-wise smokers vs Non-smokers analysis with one or more pivot table and charts:

| Count of smoker2 | Column Labels | | |
|---|---|---|---|
| **Row Labels** | **no** | **yes** | **Grand Total** |
| northeast | 79.32% | 20.68% | 100.00% |
| northwest | 82.15% | 17.85% | 100.00% |
| southeast | 75.00% | 25.00% | 100.00% |
| southwest | 82.15% | 17.85% | 100.00% |
| **Grand Total** | **79.52%** | **20.48%** | **100.00%** |



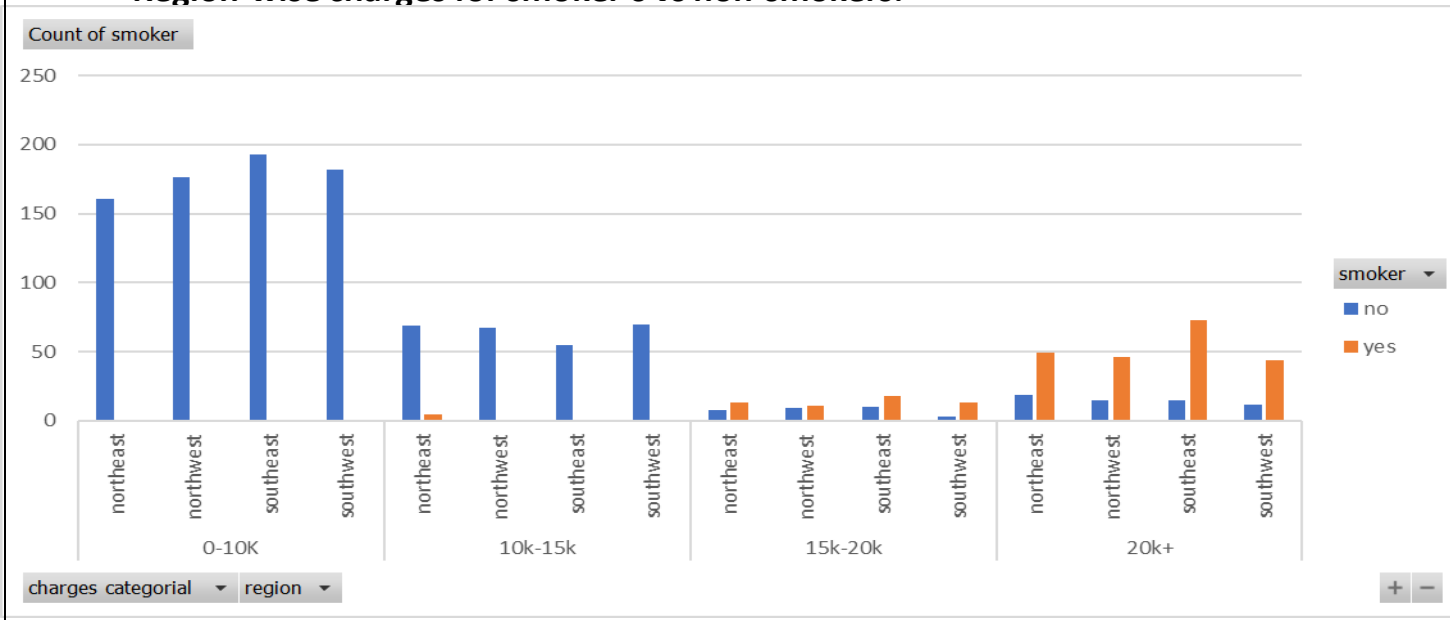

In this pivot chart we can observe that:

1) In the northeast, the smoking rate is 20%, with 2 out of 10 people being smokers.

2) Both the northwest and southwest regions exhibit a smoking prevalence of approximately 17.85%.

3) In the southeast, nearly 25% of the population engages in smoking.

**Region-wise charges for smoker's vs non-smokers:**

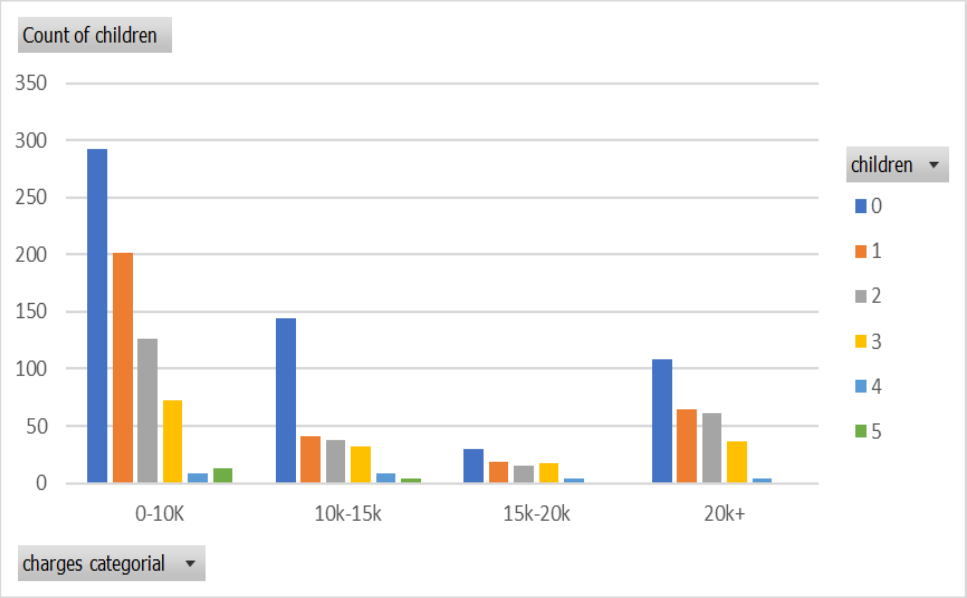



- In this dataset, it is noticeable that no smokers are billed in the insurance category of 0-10k in any region.
- In the insurance category of 0-15k, the majority of individuals who billed their medical expenses are non-smokers. There are only 7 individuals billed under the category of 10k-15k, except in the southeast region.
- In the insurance category of 15k-20k, the majority of individuals billed are smokers across all regions. However, in the southeast region, the number of claimants is somewhat higher compared to other regions.
- In the insurance category of 20+, there are approximately 273 claimants, with 212 individuals being smokers and 61 being non-smokers. Notably, in this dataset, people in the southeast region, who are smokers, have claimed the majority of medical bills, while claimant numbers in other regions are relatively similar.

| Count of smoker / Row Labels | Column Labels: no | yes | Grand Total |
|---|---|---|---|
| **0-10K** | **712** | | **712** |
| northeast | 161 | | 161 |
| northwest | 176 | | 176 |
| southeast | 193 | | 193 |
| southwest | 182 | | 182 |
| **10k-15k** | **261** | **7** | **268** |
| northeast | 69 | 5 | 74 |
| northwest | 67 | 1 | 68 |
| southeast | 55 | | 55 |
| southwest | 70 | 1 | 71 |
| **15k-20k** | **30** | **55** | **85** |
| northeast | 8 | 13 | 21 |
| northwest | 9 | 11 | 20 |
| southeast | 10 | 18 | 28 |
| southwest | 3 | 13 | 16 |
| **20k+** | **61** | **212** | **273** |
| northeast | 19 | 49 | 68 |
| northwest | 15 | 46 | 61 |
| southeast | 15 | 73 | 88 |
| southwest | 12 | 44 | 56 |
| **Grand Total** | **1064** | **274** | **1338** |

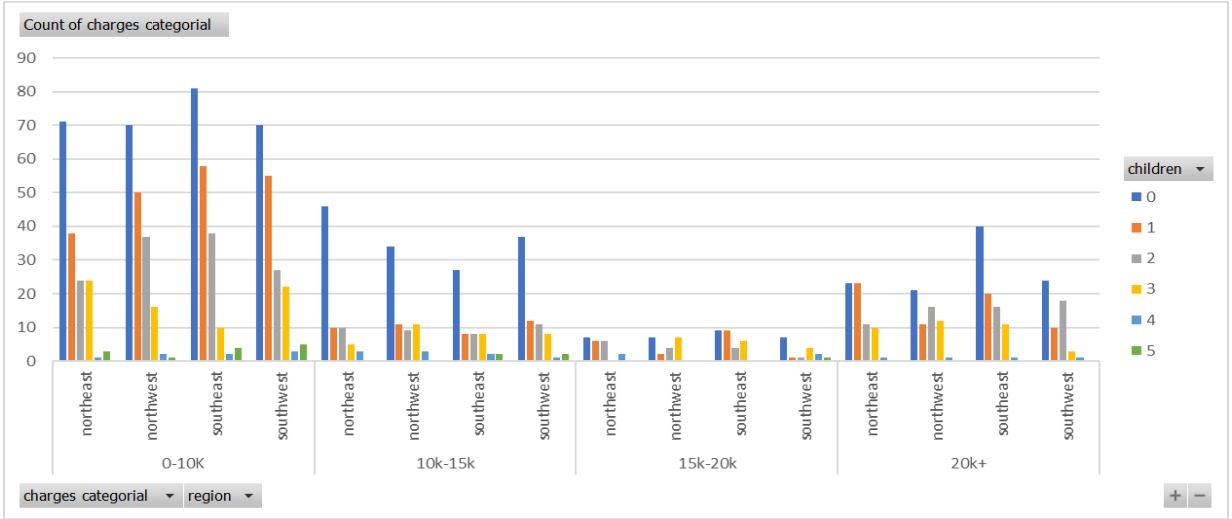
## Have charges got something to do with the number of dependents?

| Count of children | Column Labels | | | | | | |
|---|---|---|---|---|---|---|---|
| Row Labels | 0 | 1 | 2 | 3 | 4 | 5 | Grand Total |
| 0-10K | 292 | 201 | 126 | 72 | 8 | 13 | 712 |
| 10k-15k | 144 | 41 | 38 | 32 | 9 | 4 | 268 |
| 15k-20k | 30 | 18 | 15 | 17 | 4 | 1 | 85 |
| 20k+ | 108 | 64 | 61 | 36 | 4 | | 273 |
| **Grand Total** | **574** | **324** | **240** | **157** | **25** | **18** | **1338** |




- Yes, in response to the earlier question, it is confirmed that even dependents of the customer have claimed their medical bills with the insurance company.
- Specifically, individuals with 0 dependents have claimed approximately 40%, those with 1 dependent have claimed around 20%, and the remaining 40% is claimed by individuals with 2 to 5 dependents.

## Analysis of charges of dependent region-wise:

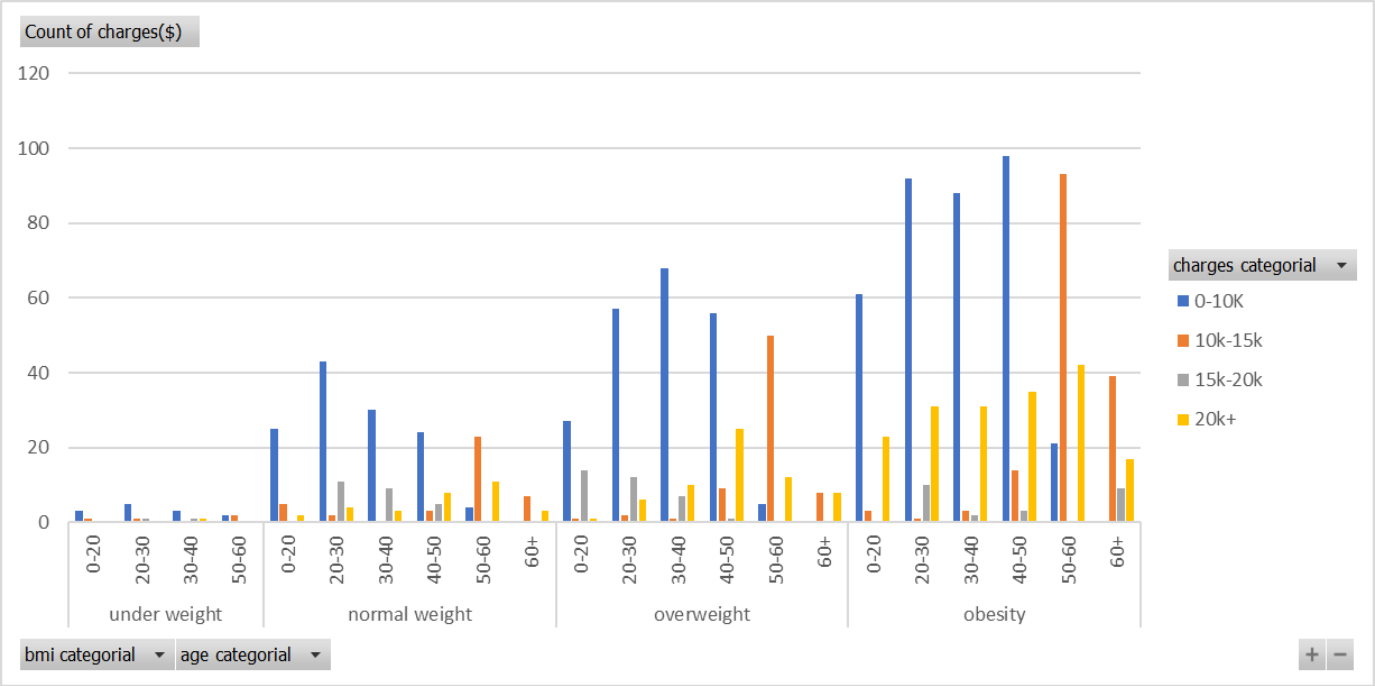


| Count of charges categorial | Column Labels | | | | | | |
|---|---|---|---|---|---|---|---|
| Row Labels | 0 | 1 | 2 | 3 | 4 | 5 | Grand Total |
| **0-10K** | **292** | **201** | **126** | **72** | **8** | **13** | **712** |
| northeast | 71 | 38 | 24 | 24 | 1 | 3 | 161 |
| northwest | 70 | 50 | 37 | 16 | 2 | 1 | 176 |
| southeast | 81 | 58 | 38 | 10 | 2 | 4 | 193 |
| southwest | 70 | 55 | 27 | 22 | 3 | 5 | 182 |
| **10k-15k** | **144** | **41** | **38** | **32** | **9** | **4** | **268** |
| northeast | 46 | 10 | 10 | 5 | 3 | | 74 |
| northwest | 34 | 11 | 9 | 11 | 3 | | 68 |
| southeast | 27 | 8 | 8 | 8 | 2 | 2 | 55 |
| southwest | 37 | 12 | 11 | 8 | 1 | 2 | 71 |
| **15k-20k** | **30** | **18** | **15** | **17** | **4** | **1** | **85** |
| northeast | 7 | 6 | 6 | | 2 | | 21 |
| northwest | 7 | 2 | 4 | 7 | | | 20 |
| southeast | 9 | 9 | 4 | 6 | | | 28 |
| southwest | 7 | 1 | 1 | 4 | 2 | 1 | 16 |
| **20k+** | **108** | **64** | **61** | **36** | **4** | | **273** |
| northeast | 23 | 23 | 11 | 10 | 1 | | 68 |
| northwest | 21 | 11 | 16 | 12 | 1 | | 61 |
| southeast | 40 | 20 | 16 | 11 | 1 | | 88 |
| southwest | 24 | 10 | 18 | 3 | 1 | | 56 |
| **Grand Total** | **574** | **324** | **240** | **157** | **25** | **18** | **1338** |

- From above data we can observe that from all region dependents have claimed their medical bill where every region is similar in number of individual (between 325) except southeast with bit higher individual (364).
- Most of dependent have claimed their bill in the insurance category of 0-10k from every region.
- In the insurance category of 10k-15k, every individual with dependents has claimed their bill, except for those in the northwest and northeast regions who have 5 dependents.
- In the insurance category of 15k-20k, there are no individuals with 3 dependents from the northeast, and no individuals with 4 or 5 dependents from the northeast, northwest and southeast regions have claimed their medical bills.
- In the insurance category of 20k+, no individuals with 5 dependents have claimed medical bills.

## Pivot table and chart of my own choice with explanation:

Data analysis of weight category, insurance charge category and age category.




- In this dataset, most people in the obesity weight category have claimed their medical bills, compared to those in the underweight and other weight categories.
- Here most of the individual aged between 40-60 have claimed their bill in the insurance category of 20k+

From this dataset, it can be concluded that individuals in the obesity category, particularly those aged between 40-60+, are claiming more medical bills compared to individuals in other age categories.

| Count of charges($) | Column Labels | | | | |
|---|---|---|---|---|---|
| Row Labels | 0-10K | 10k-15k | 15k-20k | 20k+ | Grand Total |
| **under weight** | **13** | **4** | **2** | **1** | **20** |
| 0-20 | 3 | 1 | | | 4 |
| 20-30 | 5 | 1 | 1 | | 7 |
| 30-40 | 3 | | 1 | 1 | 5 |
| 50-60 | 2 | 2 | | | 4 |
| **normal weight** | **126** | **40** | **25** | **31** | **222** |
| 0-20 | 25 | 5 | | 2 | 32 |
| 20-30 | 43 | 2 | 11 | 4 | 60 |
| 30-40 | 30 | | 9 | 3 | 42 |
| 40-50 | 24 | 3 | 5 | 8 | 40 |
| 50-60 | 4 | 23 | | 11 | 38 |
| 60+ | | 7 | | 3 | 10 |
| **overweight** | **213** | **71** | **34** | **62** | **380** |
| 0-20 | 27 | 1 | 14 | 1 | 43 |
| 20-30 | 57 | 2 | 12 | 6 | 77 |
| 30-40 | 68 | 1 | 7 | 10 | 86 |
| 40-50 | 56 | 9 | 1 | 25 | 91 |
| 50-60 | 5 | 50 | | 12 | 67 |
| 60+ | | 8 | | 8 | 16 |
| **obesity** | **360** | **153** | **24** | **179** | **716** |
| 0-20 | 61 | 3 | | 23 | 87 |
| 20-30 | 92 | 1 | 10 | 31 | 134 |
| 30-40 | 88 | 3 | 2 | 31 | 124 |
| 40-50 | 98 | 14 | 3 | 35 | 150 |
| 50-60 | 21 | 93 | | 42 | 156 |
| 60+ | | 39 | 9 | 17 | 65 |
| **Grand Total** | **712** | **268** | **85** | **273** | **1338** |

## Inferences on descriptive statistics of given variable:

| | *age* | *children* | *bmi* | *smoker* | *northwest* | *Southeast* | *southwest* | *charges($)* | *gender categorial* |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | |
| Mean | 39.20702541 | 1.094917788 | 30.66339686 | 0.204783259 | 0.242899851 | 0.272047833 | 0.242899851 | 13270.42227 | 0.505231689 |
| Standard Error | 0.384102419 | 0.032956155 | 0.166714232 | 0.01103632 | 0.011728017 | 0.012170498 | 0.011728017 | 331.0674543 | 0.013673526 |
| Median | 39 | 1 | 30.4 | 0 | 0 | 0 | 0 | 9382.033 | 1 |
| Mode | 18 | 0 | 32.3 | 0 | 0 | 0 | 0 | 1639.5631 | 1 |
| Standard Deviation | 14.04996038 | 1.20549274 | 6.098186912 | 0.403694038 | 0.428995407 | 0.445180784 | 0.428995407 | 12110.01124 | 0.500159569 |
| Sample Variance | 197.4013867 | 1.453212746 | 37.18788361 | 0.162968876 | 0.18403706 | 0.19818593 | 0.18403706 | 146652372.2 | 0.250159595 |
| Kurtosis | -1.245087653 | 0.202454147 | -0.050731531 | 0.145755539 | -0.559856699 | -0.949522817 | -0.559856699 | 1.606298653 | -2.002556636 |
| Skewness | 0.055672516 | 0.93838044 | 0.284047111 | 1.46476616 | 1.200409261 | 1.025621147 | 1.200409261 | 1.515879658 | -0.020951397 |
| Range | 46 | 5 | 37.17 | 1 | 1 | 1 | 1 | 62648.55411 | 1 |
| Minimum | 18 | 0 | 15.96 | 0 | 0 | 0 | 0 | 1121.8739 | 0 |
| Maximum | 64 | 5 | 53.13 | 1 | 1 | 1 | 1 | 63770.42801 | 1 |
| Sum | 52459 | 1465 | 41027.625 | 274 | 325 | 364 | 325 | 17755824.99 | 676 |
| Count | 1338 | 1338 | 1338 | 1338 | 1338 | 1338 | 1338 | 1338 | 1338 |

In the above descriptive analysis, we can extract data as follows:

- The average age of claimants is 39 years old, and the most frequently occurring age among claimants is 18. With negative kurtosis and positive skewness in the age distribution, this suggests that there is a concentration of higher-aged individuals who have claimed more from the insurance, contributing to a distribution with lighter tails and a skew towards the higher age values.
  Minimum age of claimants are 18 years and maximum are 64 years.
- The average number of dependents is 1 and most of the individual has 0 dependent. The maximum number of claimants are 5 dependents.
- The average individual having BMI of 30.66 have claimed their insurance charges. Here kurtosis if negative and skewness is positive which refers that higher the rate of BMI higher the rate of insurance claimed in higher insurance category.
- Out of 10 people 2 people are smokers, here both kurtosis and skewness are positive which explains when there is increase in smokers the claimants of the medical bill also increase.

Based on this analysis, it can be inferred that factors such as age, BMI, and smoking habits play a pivotal role in influencing the insurance company's payments for medical bills. These factors appear to contribute to financial losses for the insurance company.

## INFERENCE OF FINAL REGRESSION MODEL:

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.865849023 |
| R Square | 0.74969453 |
| Adjusted R Square | 0.748943426 |
| Standard Error | 6067.787249 |
| Observations | 1338 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 4 | 1.46996E+11 | 36748942863 | 998.1232235 | 0 |
| Residual | 1333 | 49078450117 | 36818042.1 | | |
| Total | 1337 | 1.96074E+11 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | -12102.76936 | 941.9839411 | -12.84816952 | 1.05165E-35 | -13950.70186 | -10254.83687 | -13950.70186 | -10254.83687 |
| age | 257.8495073 | 11.89638633 | 21.67460774 | 1.74834E-89 | 234.5118282 | 281.1871863 | 234.5118282 | 281.1871863 |
| children | 473.5023156 | 137.7916715 | 3.436363827 | 0.000607716 | 203.1901623 | 743.8144689 | 203.1901623 | 743.8144689 |
| bmi | 321.8514025 | 27.37763213 | 11.75599851 | 1.97399E-30 | 268.1434634 | 375.5593415 | 268.1434634 | 375.5593415 |
| smoker | 23811.39984 | 411.2197148 | 57.90432459 | 0 | 23004.69153 | 24618.10816 | 23004.69153 | 24618.10816 |

This represents the final and most effective regression model, utilizing independent variables such as age, number of children, BMI, and smoker status. With these variables 75% of dependent variable is explained.
The strength of the linear relation between dependent and independent are 86%.
When each increase in independent variable there will be increase in dependent variable.
In these independent variable smokers is having highest positive value indicating strong relationship with dependent variable.

## Conclusion:

In conclusion, this health insurance claim analysis provides valuable insights into the factors influencing insurance charges. Through comprehensive exploratory data analysis (EDA) and multiple linear regression, key variables such as age, BMI, and smoking status have been identified as significant contributors to medical charges.
The transformation of data using dummy variables enhances the predictive accuracy of the analysis. These findings offer insurers a nuanced

understanding of the determinants of insurance charges, enabling them to refine pricing strategies and make informed decisions.

This analysis serves as a foundation for data-driven decision-making in the insurance domain, empowering stakeholders to navigate complexities and optimize their approach to insurance premium determinations.