A

PROJECT REPORT

In the course PS04CSTA24 at

M.Sc. (Statistics)

To fulfilment of the Degree in M.Sc. (Statistics)

Entitled

“Statistical Insights into Health Insurance Claims and Premiums"

By

Miss. Kalyani S. Patil

M.Sc. STATISTICS (SEMESTER IV)

Under the Guidance of

Dr. Jyoti. M. Divecha

Dr. Dharmesh Raykundaliya

Dr. Khimaya Tinani

DEPARTMENT OF STATISTICS

SARDAR PATEL UNIVERSITY

VALLABH VIDYANAGAR

2024-2025

ACKNOWLEDGEMENT

Through this opportunity, I would like to thank many people who have contributed their time and knowledge to this project.

I express my deep sense of gratitude and indebtedness to Dr. Jyoti M. Divecha, Head of the Department of Statistics, for extending her help and support.

I would like to extend my sincerest gratitude to Dr. Dharmesh Raykundaliya and Dr. Khimaya Tinani for their invaluable guidance, unwavering support, and insightful contributions. Their expertise and guidance were instrumental in steering us through the project, and I owe much of its success to them.

I also thank the faculty of the department for their support. Additionally, I would like to express my gratitude to our friends for providing us with amenities and cooperation throughout the project.

Last but not least, special thanks to everyone involved in my project for providing with their information, precious time, and continued support.

Kalyani S. Patil (STAT 06)

Place: V.V Nagar, Anand

Date:

CERTIFICATE

This is to certify that **Miss. Kalyani Sharad Patil**, student ID Number:**2023014744** of Semester-IV, "**Master of Science in Statistics** Sardar Patel University, Vallabh Vidyanagar, Anand has successfully completed her project work on topic

"Statistical Insights into Health Insurance Claims and Premiums"

For paper code PS04CST54 and term ending in March -2025

Date:

Prof (Dr.) Jyoti M. Divecha	**Dr. Dharmesh Raykundaliya**	**Prof (Dr.) Khimiya Tinani**
(Head of Department)	**(Project Guide)**	**(Project Guide)**
&		
(Project Guide)		

INDEX

INTRODUCTION

The health insurance industry plays a crucial role in protecting individuals and families from unexpected medical expenses. Insurers must accurately assess risks and determine fair premium pricing to ensure financial sustainability while keeping policies affordable for customers.

One of the biggest challenges in the insurance sector is predicting the likelihood of a claim, its severity, and the risk profile of policyholders. To address this, statistical models and data analytics are widely used in insurance risk management.

This project, titled ***"Statistical Insights into Health Insurance Claims and Premiums,"*** focuses on analysing health insurance claim data to extract meaningful insights. By applying Generalized Linear Models (GLMs) and Multiple Logistic Regression, this study aims to provide a structured approach to:

Risk Prediction – Identifying key factors that influence claim frequency and claim amount.
Premium Calculation – Developing pricing strategies based on risk assessment.
Health Risk Profiling – Understanding the health status of policyholders and its impact on claims.
Fraud Detection – Recognizing patterns that indicate suspicious or fraudulent claims.
Cost Forecasting – Estimating future claim amounts based on policyholder demographics and health conditions.

By leveraging statistical techniques, this study will help insurance companies in making data-driven decisions, optimizing resources, and reducing financial risks associated with unpredictable claims.

Background of Study

Health insurance is an essential component of financial security, but it comes with several challenges for insurers. The ability to predict claim patterns, assess risks, and design fair premium structures is key to maintaining the balance between policyholder benefits and the profitability of the insurance company.

- The increasing availability of large-scale health insurance data presents an opportunity to develop accurate predictive models. These models help in:
- Identifying high-risk policyholders who are more likely to file claims.
- Classifying claims based on severity to better allocate resources.
- Reducing fraudulent activities by detecting unusual claim patterns.
- Developing risk-based pricing to ensure fair premium distribution.

"By analysing health data, we proved smoking/BMI/age drive insurance claims. Built models to set fair premiums, detect fraud (100% accuracy!), and help insurers save costs. A perfect mix of stats + real-world impact!"

OBJECTIVE

1. Identify which risk factors are most important in predict likelihood of claim to help actuaries.
2. Determine a **base premium** and adjust it according to an individual's risk category (**Low, Medium, High**).
3. To Check the Association between the variables by welch two sample t test
4. To handle under dispersion in a skewed response variable using a quasi-likelihood approach and ensure reliable modelling with a constant coefficient of variation**.**
5. Recognizing patterns that indicate suspicious or fraudulent claims.

DATA

Case 1: - Multiple Logistic Regression model

The dataset, sourced from Kaggle, contains information on policyholders and their insurance Claims The data consists of 1338 rows and 8 columns, with each row representing an individual policyholder.

➢ Data sample: -

age	sex	bmi	children	smoker	region	charges	Insurance claim
19	0	27.9	0	1	3	16884.92	1
18	1	33.77	1	0	2	1725.552	1
28	1	33	3	0	2	4449.462	0
33	1	22.705	0	0	1	21984.47	0
32	1	28.88	0	0	1	3866.855	1
31	0	25.74	0	0	2	3756.622	0
46	0	33.44	1	0	2	8240.59	1
37	0	27.74	3	0	1	7281.506	0
37	1	29.83	2	0	0	6406.411	0
60	0	25.84	0	0	1	28923.14	0
25	1	26.22	0	0	0	2721.321	1
62	0	26.29	0	1	2	27808.73	1
23	1	34.4	0	0	3	1826.843	1
56	0	39.82	0	0	2	11090.72	1
27	1	42.13	0	1	2	39611.76	1
19	1	24.6	1	0	3	1837.237	0
52	0	30.78	1	0	0	10797.34	1
23	1	23.845	0	0	0	2395.172	0
56	1	40.3	0	0	3	10602.39	1

➢ variables in the dataset: -

age	Numeric	age of policyholder
sex	Categorical	gender of policy holder (female=0, male=1)
bmi	Numeric	Body mass index
children	Numeric	number of children / no of dependents of policyholder
smoker	Categorical	smoking state of policyholder (non-smoker=0, smoker=1)
region	Categorical	the residential area of policyholder in the US (northeast=0, northwest=1, southeast=2, southwest=3)
charges	Numeric	individual medical costs billed by health insurance
Insurance claim	Categorical	yes=1, no=0

➢ Summary Statistics of numerical Attributes

	Min	1st Qu	Median	Mean	3rd Qu	Max
age	18.00	27.00	39.00	39.21	51.00	64.00
bmi	15.96	26.30	30.40	30.66	34.69	53.13
charges	1122	4740	9382	13270	16640	63770

➢ Summary Statistics of numerical Attributes

Variable	Values
Sex	0: 662, 1: 676
Children	0: 574, 1: 324, 2: 240, 3: 157, 4: 25, 5: 18
Smoker	0: 1064, 1: 274
Region	0: 324, 1: 325, 2: 364, 3: 325
Insurance Claim	0: 555, 1: 783

Case 2: -Gamma GLM Model

We have taken this dataset from Kaggle. The dataset used for this analysis contains information on policyholders and their insurance claims. The dataset contains 13650 observation and 13 variables, the following variables are available in the dataset.

age	sex	weight	bmi	hereditary_diseases	no_of_dependents	smoker	city	bloodpressure	diabetes	regular_ex	job_title	claim
60	male	64	24.3	NoDisease	1	0	NewYork	72	0	0	Actor	13112.6
49	female	75	22.6	NoDisease	1	0	Boston	78	1	1	Engineer	9567
32	female	64	17.8	Epilepsy	2	1	Phildelphia	88	1	1	Academici	32734.2
61	female	53	36.4	NoDisease	1	1	Pittsburg	72	1	0	Chef	48517.6
19	female	50	20.6	NoDisease	0	0	Buffalo	82	1	0	HomeMak	1731.7
42	female	89	37.9	NoDisease	0	0	AtlanticCit	78	0	0	Dancer	6474
18	male	59	23.8	NoDisease	0	0	Portland	64	0		Singer	1705.6
21	male	52	26.8	NoDisease	0	0	Cambridge	74	1	0	Actor	1534.3
40	female	69	29.6	NoDisease	0	0	Springfield	64	1	1	DataScient	5910.9
51	female	50	33	EyeDisease	0	1	Syracuse	0	1	0	Police	44400.4
59	female	68	36.5	NoDisease	1	0	Baltimore	70	1	1	HomeMak	28287.9
19	male	45	24.6	NoDisease	1	0	York	0	0	1	Student	1837.2
21	female	53	35.7	NoDisease	0	0	Trenton	62	1	0	Singer	2404.7
27	male	53	18.9	NoDisease	3	0	Warwick	90	1	0	Doctor	4827.9
56	male	67	40.3	NoDisease	0	0	Washingtc	0	1	0	Engineer	10602.4
56	female	69	27.2	NoDisease	0	0	Providence	68	1	0	Chef	11073.2
63	male	67	41.3	NoDisease	3	0	Harrisburg	70	1	1	Singer	15555.2
19	female	46	24.6	NoDisease	1	0	Newport	70	1	0	Student	2709.2
52	female	76	38.4	NoDisease	2	0	Stamford	48	1	0	Manager	11396.9

- **variables in the dataset:-**

age	Numeric	Age of the policyholder
sex	Categorical	Gender of the policyholder
weight	Numeric	Weight of the policyholder
BMI	Numeric	Body Mass Index, It is an objective index of body weight (kg/m2), calculated using the ratio of height to weight
No of dependents	Numeric	Number of dependent persons on the policyholder
smoker	Categorical	Indicates whether the policyholder is a smoker or a non-smoker (0 = nonsmoker, 1 = smoker)
claim	Continuous	The amount claimed by the policyholder
blood pres	Numeric	Blood pressure reading of the policyholder
diabetes	Categorical	Indicates whether the policyholder suffers from diabetes (0 = non-diabetic, 1 = diabetic)
regular ex	Categorical	indicates whether the policyholder exercises regularly (0 = no exercise, 1 = exercises regularly)
job title	Character	Job profile of the policyholder
city	Character	The city in which the policyholder resides
hereditary diseases	Character	Indicates whether the policyholder suffers from hereditary diseases

➢ Summary Statistics of numerical Attributes

	Min	1st Qu	Median	Mean	3rd Qu	Max
age	18	27	40	39	52	64
weight	34	54	63	64	75	95
bmi	16	25	29.4	30.29	34.4	53.1
blood pressure	0	64	72	68.63	80	122
claim	1122	4889	9716	13416	16451	63770

Methodology

➢ For case 1

Chi Square Test

Chi-square test is a statistical test used to determine whether there is a significant association between two categorical variables. It analyses the relationship between variables that can be classified into categories. Hypotheses:

Null Hypothesis (H0): There is no association between the two categorical variables.

Alternative Hypothesis (H1): There is an association between the two categorical variables.

Pearson Chi Square

Test Pearson's chi-square test is a statistical test that compares categorical data to expected values to determine if the data is significantly different. It is also known as the chi-square test. The test also assesses three types of comparison: goodness of fit, homogeneity, and independence. The test compares observed data to a model that distributes the data according to the expectation that the variables are independent. If the observed data doesn't fit the model, the likelihood that the variables are dependent becomes stronger. The test then computes a test statistic called the chi-squared statistic which measures the discrepancy between the observed and expected frequencies. The formula for the chi-squared statistic is

$$\boldsymbol{\chi^2 = \sum \frac{(0 - E)^2}{E}}$$

Where,

χ^2 is the chi-square test statistic

O is the observed frequency

E is the expected frequency

Generalized Linear Model

Generalized linear models (GLMs) relate the response variable which we want to predict, to the explanatory variables or factors about which we have information. Following are the three components which we need to specify in defining GLM:

Key Components:

1) **A distribution of the response variable:**

For linear models, the response variable had a normal distribution, Y ∼ N(0,σ)We now extend this to a general form of distributions known as the exponential family.

Exponential Family:

The exponential family is the set of distributions whose probability function, or probability density function (PDF), can be written in the following form:

$$f_y(y;\ \theta, \phi) = \exp\left[\frac{y\theta\ -\ b(\theta)}{a(\phi)} + c(y, \phi)\right]$$

Where $a(\phi)$, $b(\theta)$and $c(y, \phi)$ are specific functions:

- $b(\theta)$ is a function that depends only on θ.
- $a(\phi)$ is a function that depends only on ϕ
- $c(y,\phi)$ is a function that depends on ϕ and y but not on θ.

The expected value and variance are given by

$$E(Y\ ) = b'(\theta)$$
$$Var(Y\ ) = a(\phi)b''(\theta)$$

There are two parameters in the above PDF: θ, which is called the 'natural' parameter and is relevant to the model for relating the response (Y) to the covariates, and ϕ, known as the scale parameter or dispersion parameter.

2) **A Linear Predictor**

The linear predictor, η, is a function of the covariates. For the bivariate linear regression model, this was $\beta 0 + \beta 1x$. For the multivariate linear regression model, this was $\beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta kxk$, which we then extended to functions of the explanatory variables. There are two kinds of covariates used in GLMs: variables and factors.

Variables

In general, variables are covariates where the actual value of a variable enters the linear predictor. The age of the policyholder is an actuarial example of a variable.

Factors

The other main type of covariate is a factor, which takes a categorical value. For example, the sex of the policyholder is either male or female, which constitutes a factor with 2 categories (or levels).

3) A Link Function

The link function connects the mean response to the linear predictor, $g(\mu) = \eta$, where $\mu = E(Y)$ For linear models, the mean response was equal to the linear predictor, e.g., $\mu = E(Y) = \beta 0 + \beta 1x$, so the link function is the identity function: $g(\mu) = \mu$. The link function is the connection between the linear predictor (input) and the mean of the distribution (output).

The link function connects the mean response to the linear predictor: $g(\mu) = \eta$, where

$$E(Y) = \mu.$$

Technically, it is necessary for the link function to be differentiable and invertible in order to fit a model.

MULTIPLE LOGISTIC REGRESSION

Like ordinary regression, logistic regression extends to models with multiple explanatory variables. For instance, the model for

$\pi(x) = P(Y = 1)$ at values $x = (x_1, \ldots., x_p)$ of p predictors is

$$logit[\pi(x)] = \propto + \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \ldots\ldots + \beta_p X_p$$

The alternative formula, directly specifying $\pi(x)$ is

$$\pi(x) = \frac{\exp(\propto + \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}{1 + \exp(\propto + \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}$$

The parameter β_i refers to the effect of x_i on the log odds that Y=1, controlling the other x_i For instance, exp (β_i) is the multiplicative effect on the odds of a 1-unit increase in x_i, at fixed levels of other x_j.

An explanatory variable can be qualitative, using dummy variables for categories.

Odds ratio

For a probability π of success, the odds are defined to be

$$\Omega = \frac{\pi}{1-\pi}$$

The odds are nonnegative, with $\Omega > 1.0$ when a success is more likely than a failure. When $\pi = 0.75$,for instance,

then $\Omega = \frac{0.75}{1-0.75} = 3.0$; a success is three times as likely as a failure, and we expect about three successes for 1 every one failure. When $\Omega = \frac{1}{3}$ a failure is three times as likely as a success. Inversely,

$$\pi = \Omega/(\Omega + 1)$$

For instance, when $\Omega = \frac{1}{3}$,then π =0.25.

Refer again to a 2*2 table. Within row *i,* the odds of success instead of failure are $\Omega_i = \frac{\pi_i}{1-\pi_i}$.

The ratio of the odds and Ω_1 and Ω_2 in the two rows,

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_{2)}}$$

is called the **odds ratio.**

- **For case 2: -**

Welch T-Test: The **Welch T-Test** is a statistical test used to compare the means of two independent groups when their variances are unequal. It is a modification of the standard **Student's T-Test**, which assumes equal variances. Welch's T-Test is more reliable when sample sizes and variances differ between the two groups.

The test statistic for Welch's T-Test is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Hypotheses

- **Null Hypothesis (H_0)**: The means of the two groups are equal ($\mu_1 = \mu_2$)
- **Alternative Hypothesis (H_1)**: The means of the two groups are different ($\mu_1 \neq \mu_2$)

When to Use Welch's T-Test?

- When comparing two independent groups
- When variances are unequal (heteroscedasticity)
- When sample sizes are different

MODEL WITH CONSTANT CV

Myers and Montgomery (1997) Like the exponential distribution, the gamma distribution also finds applications in inter-arrival time problems. In addition, the gamma distribution has potential applications in regression problems in which the response is continuous and the variance is not constant but rather is proportional to the square of the mean. Such a condition implies a constant coefficient of variation. There are other alternatives to the use of the gamma distribution in this case.

One possible option is to use a natural log transformation on the response, which stabilizes the variance. In the case of this transformation, all coefficients are unbiased except the intercept. This approach inherently assumes that the /mean distribution of the response is log-normal. The intercept is biased by $(\sigma/\mu)^2/2$, since, from a Taylor series expansion, we know that

$$E[\ln(y)] = \ln\mu - \frac{(\sigma/\mu)^2}{2}$$

A second approach assumes a gamma distribution and appeals to the framework of GLMs.

Consider the density function for the gamma distribution, which is

$$f_x(x,a,s) = \frac{1}{s^{\alpha}\Gamma\alpha} x^{\alpha-1} e^{\frac{x}{s}} \quad ; x > 0, a > 0, s > 0$$

Were,

α is shape parameter

s is scale parameter

so, the $E(x) = \frac{\alpha}{\lambda}$ and $v(x) = \frac{\alpha}{\lambda^2}$

Here,

$$\text{Coeff of variation} = \frac{sd}{mean} = \frac{\sqrt{v(x)}}{E(x)}$$

If plot is positively skewed then use lognormal and gamma distribution.

OVERDISPERSION AND UNDERDISPERTION

Overdispersion is an important concept in the analysis of discrete data. Many times, data admit more variability than expected under the assumed distribution. The extra variability not predicted by the generalized linear model random component reflects overdispersion. Overdispersion occurs because the mean and variance components of a GLM are related and depend on the same parameter that is being predicted through the predictor set.

For the binomial response, if $Yi \sim Bin(ni, \pi i)$, the mean is $\mu i = ni\pi i$, and the variance is $\mu i(ni - \mu i)/ni$.

- **Overdispersion** means that the variance of the response Yi is greater than what's assumed by the model.
- **Under dispersion** is also theoretically possible but rare in practice. More often than not, if the model's variance doesn't match what's observed in the response, it's because the latter is greater

Adjusting for Overdispersion

The most popular method for adjusting for overdispersion comes from the theory of quasi-likelihood. Quasi likelihood has come to play a very important role in modern statistics. It is the foundation of many methods that are thought to be "robust" (e.g. Generalized Estimating Equations (GEE) for longitudinal data) because they do not require the specification of a full parametric model. In the quasi likelihood approach, we must first specify the "mean function" which determines how $\mu_i = E(Y_i)$ is related to the covariates. In the context of logistic regression, the mean function is

$$\mu_i = n_i exp(x_i^T \beta)$$

which implies

$$log\ (\frac{\pi_i}{1 - \pi_i}) = x_i^T \beta$$

Then we must specify the "variance function," which determines the relationship between the variance of the response variable and its mean. For a binomial model, the variance function is $\mu_i(n_i - \mu_i)/n_i$. But to account for overdispersion, we will include another factor σ^2 called the "scale parameter," so that

$$V(Yi) = \sigma^2 \mu_i(n_i - \mu_i)/n_i$$

If $\sigma^2 \neq 1$ then the model is not binomial;

- σ^2>1 corresponds to "overdispersion",
- σ^2<1 corresponds to "under dispersion."

supervised machine learning (Random Forest)

Random Forest is a widely-used machine learning algorithm developed by Leo Breman and Adele Cutler, which combines the output of multiple decision trees to reach a single result**.** Its widespread popularity stems from its user-friendly nature and adaptability, enabling it to tackle both classification and regression problems effectively. The algorithm's strength lies in its ability to handle complex datasets and mitigate overfitting, making it a valuable tool for various predictive tasks in machine learning.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing *continuous variables,* as in the case of regression, and *categorical variables,* as in the case of classification. It performs better for classification and regression tasks. In this tutorial, we will understand the working of random forest and implement random forest on a classification tasks

- **Assumptions of Random Forest**

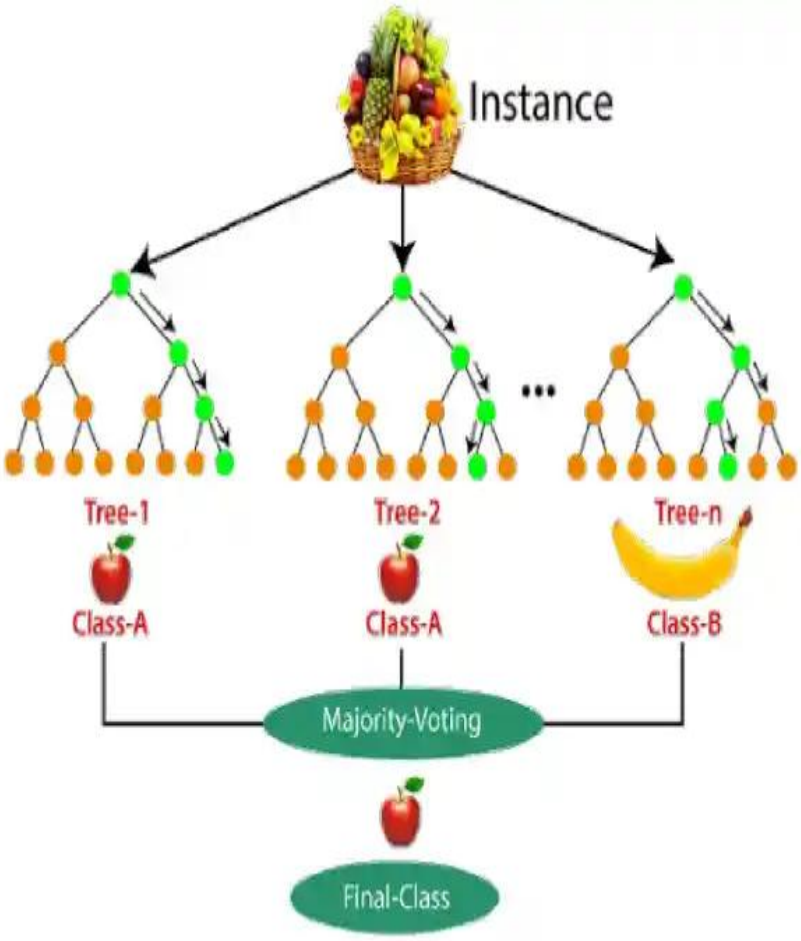
- **Independence of Trees**: The decision trees in the forest should be independent of each other. We achieve this through bootstrap sampling and feature randomness.
- **Sufficient Data**: Random Forest requires a large amount of data to build diverse trees and achieve optimal performance.
- **Balanced Trees**: The algorithm assumes that individual trees grow sufficiently deep to capture the underlying patterns in the data.
- **Noisy Data Handling**: Random Forest can handle noisy data, but it assumes that the noise randomly distributes and is not systematic

➢ Random Forest Applications

- **Customer churn prediction**: Businesses can use random forests to predict which customers are likely to churn (cancel their service) so that they can take steps to retain them. For example, a telecom company might use a random forest model to identify customers who are using their phone less frequently or who have a history of late payments.
- **Fraud detection**: Random forests can identify fraudulent transactions in real-time. For instance, a bank might employ a random forest model to spot transactions made from unusual locations or involving unusually large amounts of money.
- **Stock price** prediction: It can predict future stock prices. However, it is important to note that stock price prediction is a very difficult task, and no model is ever going to be perfectly accurate.

➢ For example:

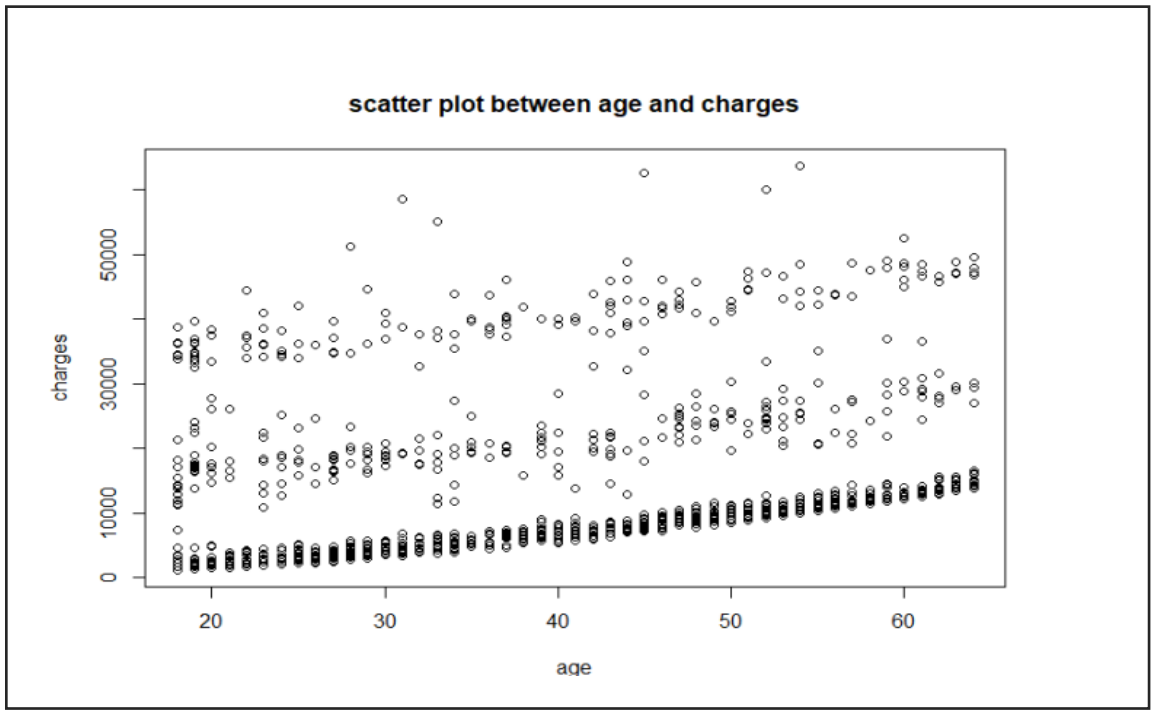
- Consider the fruit basket as the data as shown in the figure below. Now n number of samples are taken from the fruit basket, and an individual decision tree is constructed for each sample. Each decision tree will generate an output, as shown in the figure. The final output based on majority voting. In the figure below, you can see that the majority decision tree outputs an apple compared to a banana, so we take the final output as an apple.



Explanatory Data Analysis

For case 1:

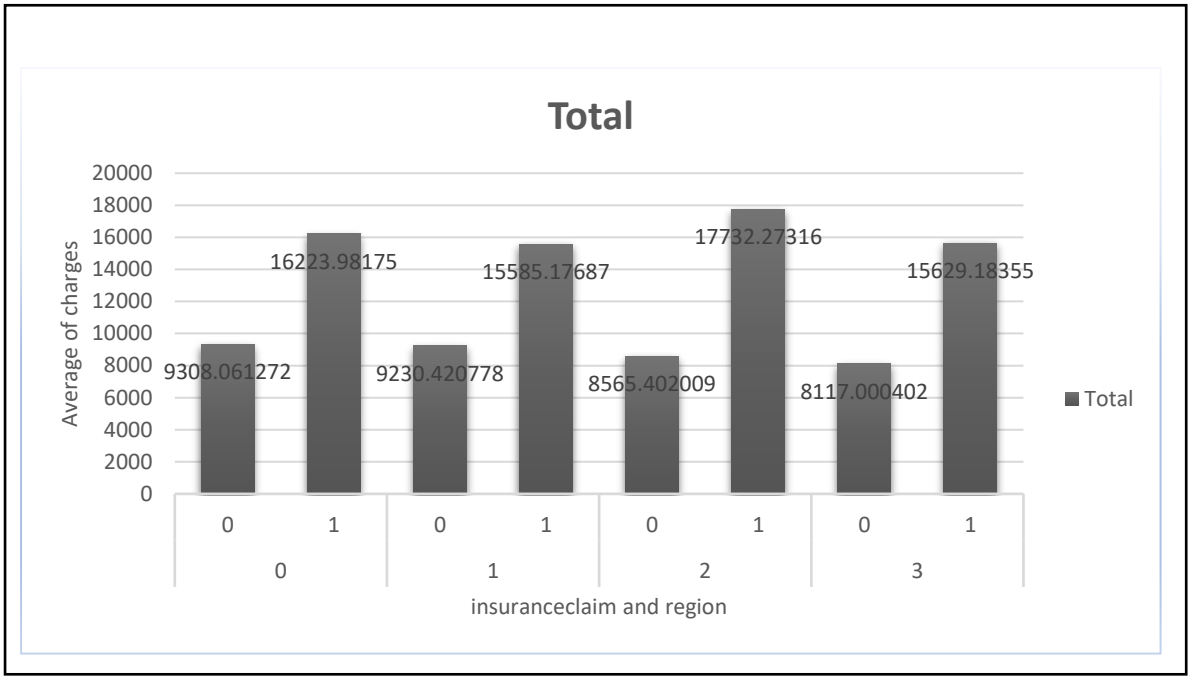
❑ Scatter Plot Between Age and Charges



Interpretation:

From the scatter plot we can see that there is strong positive relation between Age and Charges. Hence, Age has Significant effect on Insurance Charges.

❑ Average Insurance Charges by Claim Status and Region



Interpretation:

In all regions, those who filed a claim (1) had higher average charges than those who did not file a claim (0).

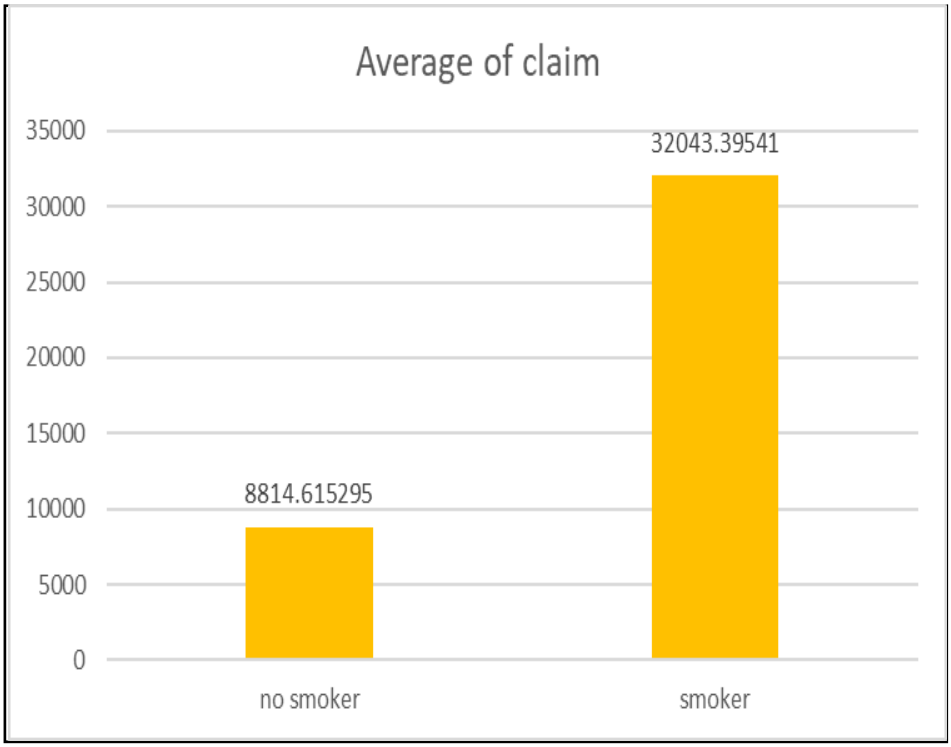
The highest average charge is observed in Region 2 for claimants (1), around 17,732, while the lowest average charge is in Region 3 for non-claimants (0), around 8,117.

This trend suggests that individuals who file insurance claims tend to have higher medical costs, regardless of the region

For case 2

❑ Graphs on impact of smoking status, regular exercise and diabetes on average of claims

- **Smokers vs. Non-Smokers:**

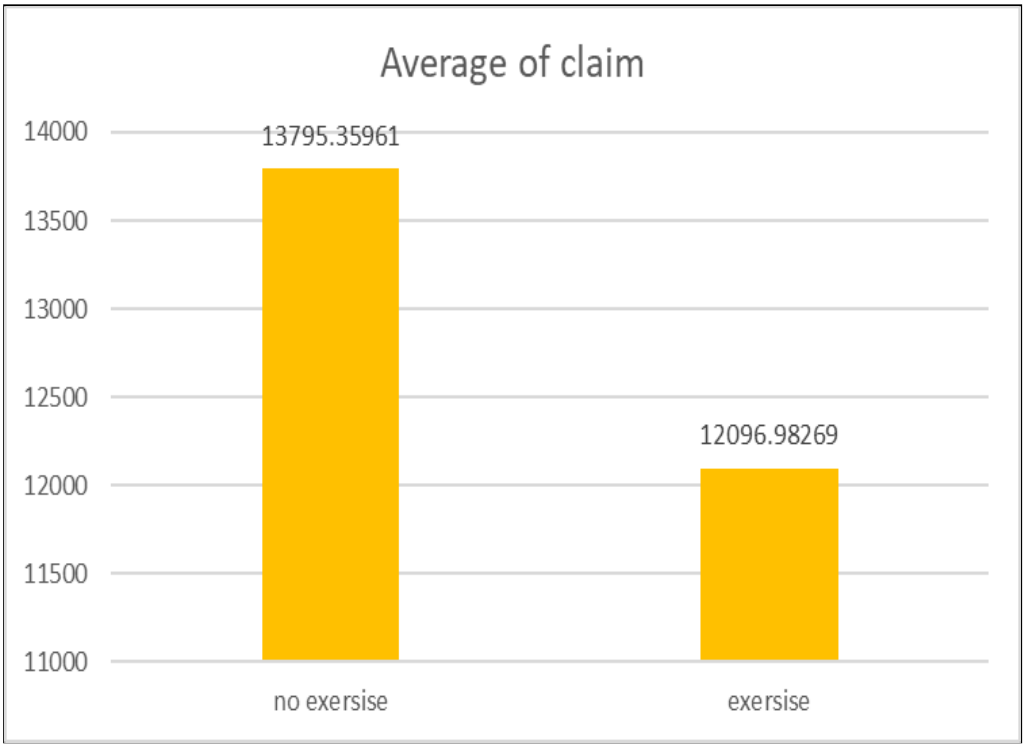


Interpretation:

The average claim amount for **smokers (32,043.40)** is significantly higher than for **non-smokers (8,814.62)**.

This suggests that smoking is associated with **higher medical costs**, possibly due to smoking-related health conditions.

- **Exercise vs. No Exercise:**

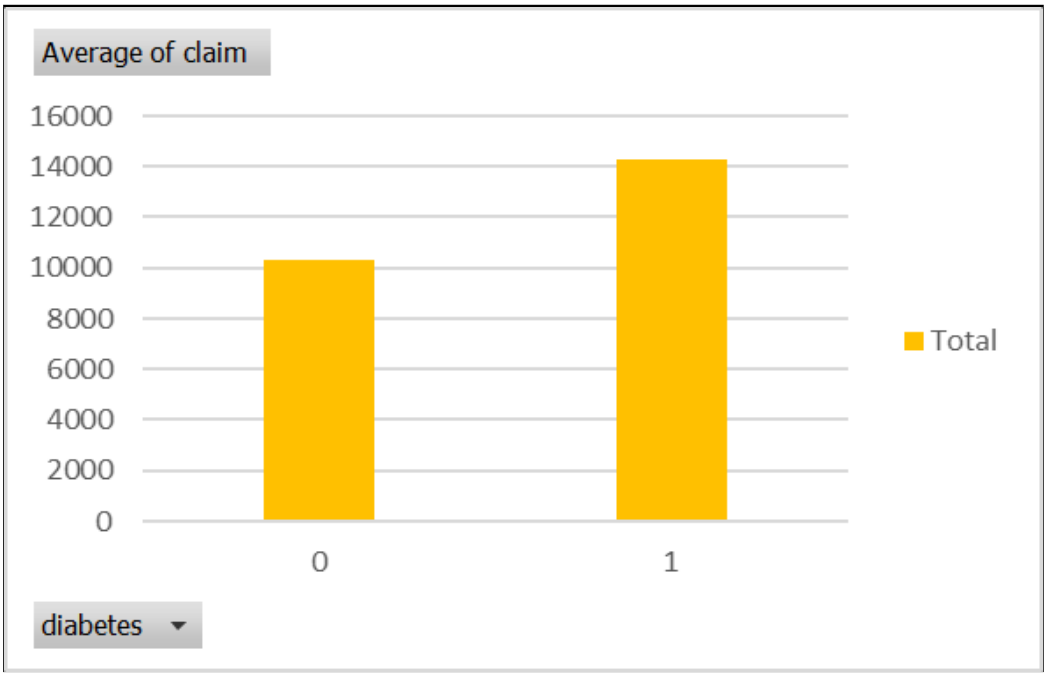


Interpretation:

People who do **not exercise (13,795.36)** have a higher average claim amount than those who **exercise regularly (12,096.98)**.

This trend indicates that a **lack of physical activity may contribute to higher medical costs**, possibly due to a higher risk of chronic diseases.

- **Diabetes Status (0 = No, 1 = Yes):**

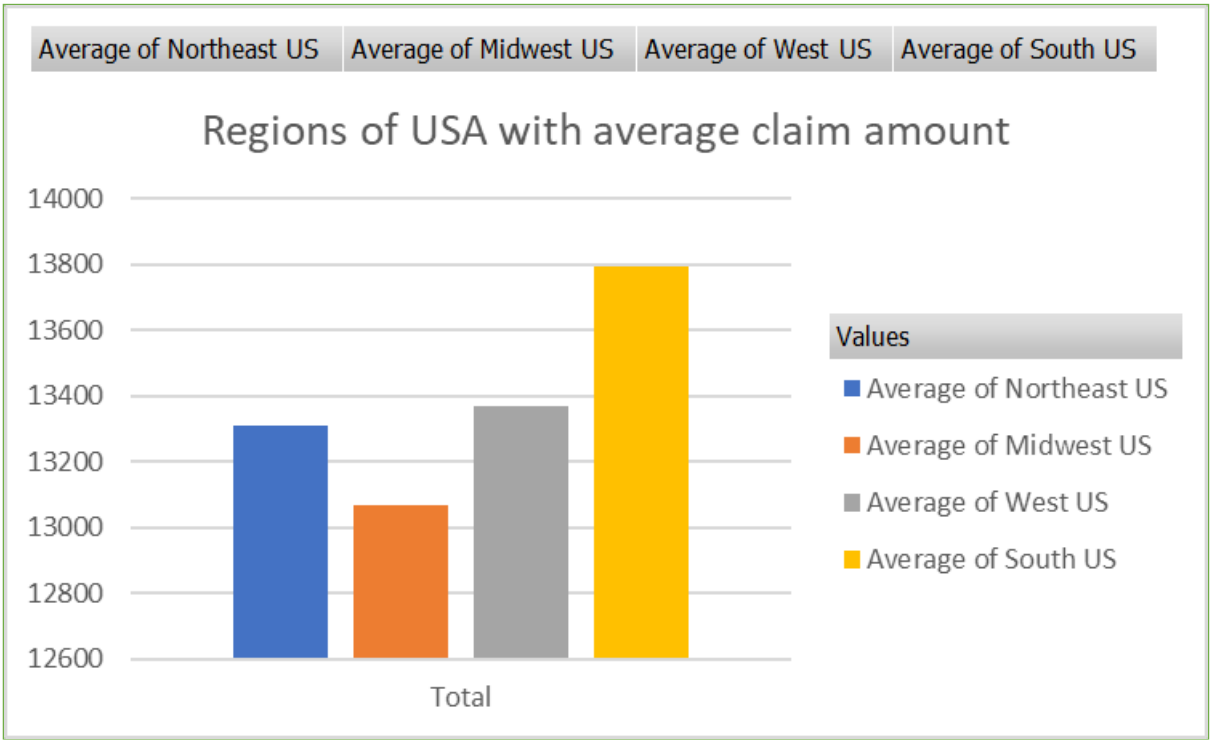


Interpretation:

The average claim amount for **diabetic patients** is higher than for **non-diabetics**

This supports the idea that diabetes leads to increased medical expenses, including medication, treatments, and complications.

❑ Graphs on Regions average claim amount

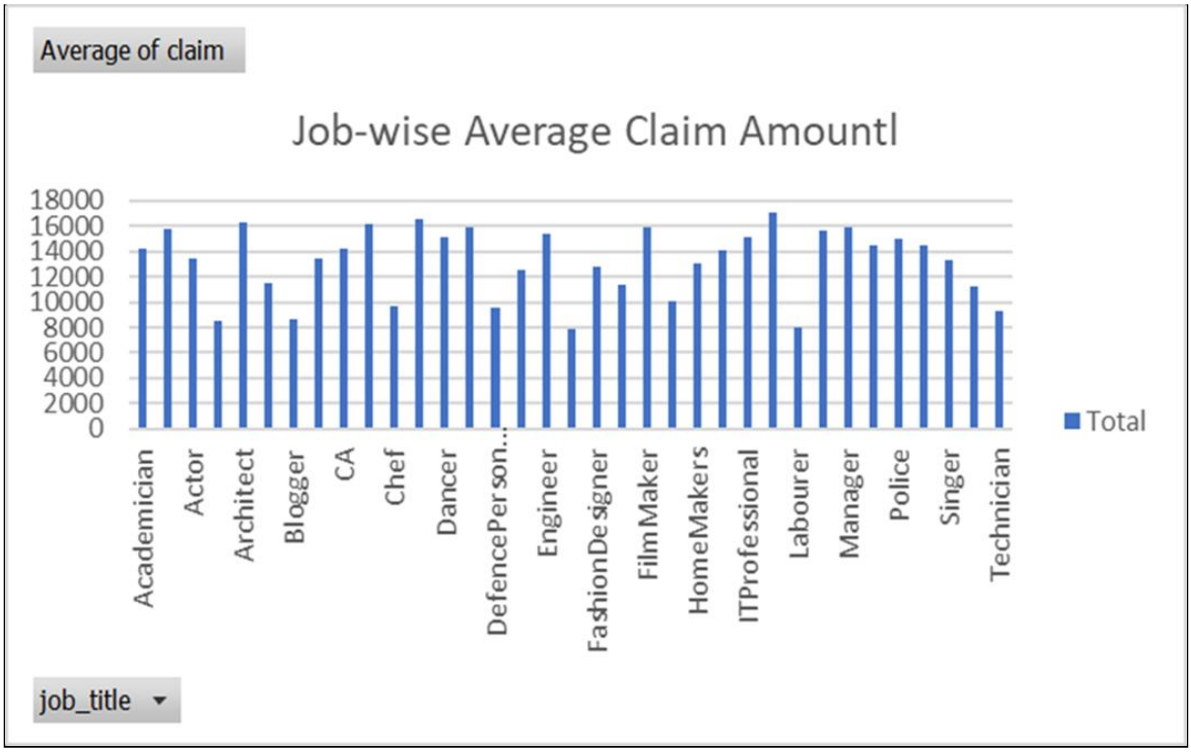


Average of Northeast US	Average of Midwest US	Average of West US	Average of South US
13310.18872	13069.32385	13366.59001	13793.60253

Interpretation:

- The South US region has the highest average claim amount (~13,800), indicating that healthcare costs or medical needs in this region might be higher.
- The South may have higher obesity rates, smoking prevalence, and chronic diseases (like diabetes and heart disease), leading to increased medical expenses**.**
- The Midwest US has the lowest average claim amount (~13,000), suggesting relatively lower medical expenses
- The Midwest, with healthier lifestyle habits and better preventive care, might experience lower claims.

❑ Job-wise Average Claim Amounts



Interpretation:

High-Risk Jobs (Higher Claims):

- Defence personnel, Police, and Labourers may have physically demanding jobs, leading to a higher risk of injuries and medical expenses.
- Managers might experience stress-related health issues (e.g., hypertension, diabetes) leading to increased claims.

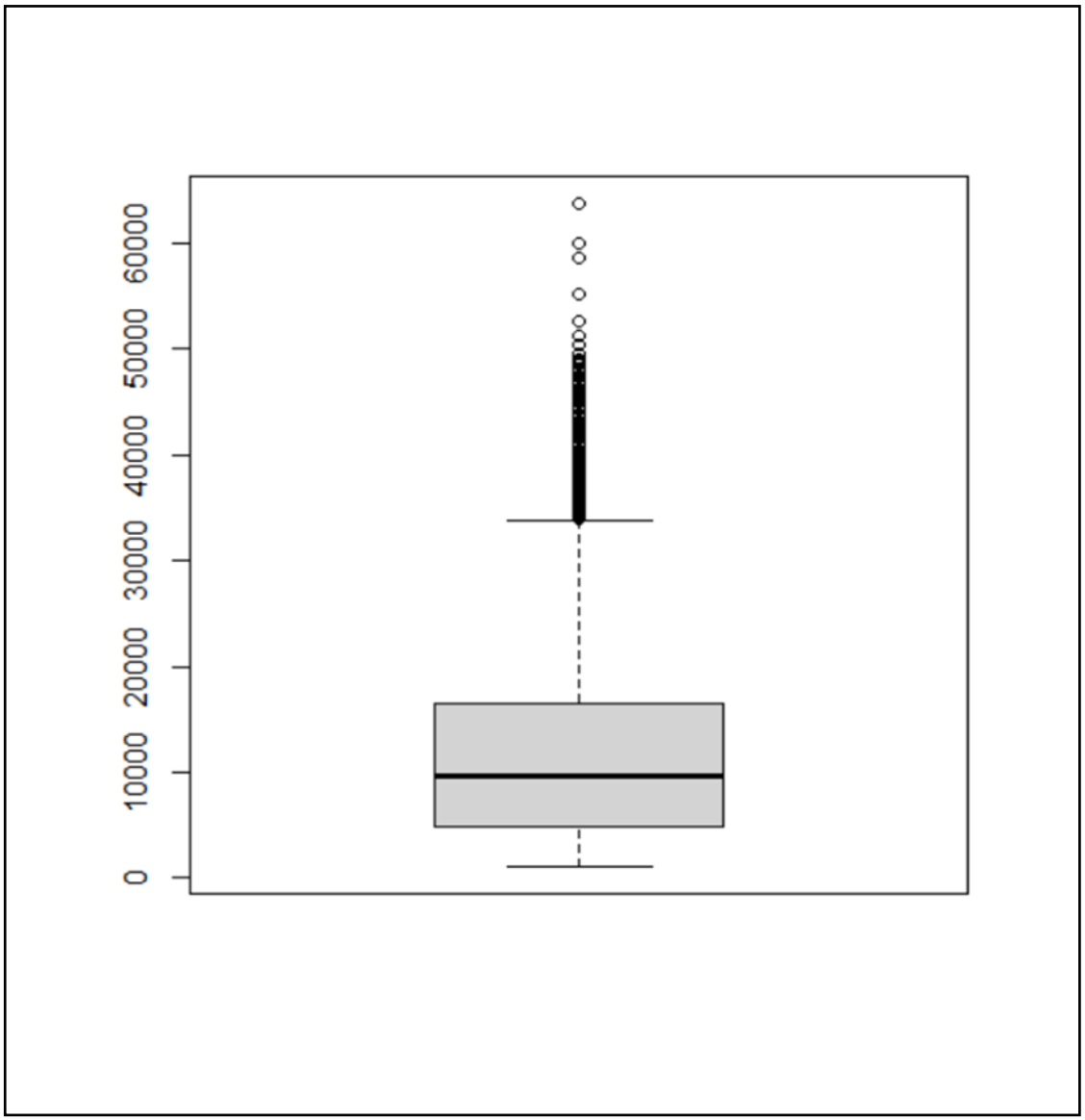
Sedentary Jobs (Moderate to Lower Claims):

- IT Professionals, Bloggers, and Academicians often have desk jobs, leading to lower accident risks but possibly higher rates of chronic illnesses over time.

Creative & Freelance Jobs (Variable Claims):

- Actors, Fashion Designers, and Filmmakers may have varying access to health insurance depending on their employment structure, affecting claim amounts.

❑ Boxplot of response variable



Interpretation:

- The distribution of the response variable is likely to be positively skewed
- The Lognormal or Gamma distribution is appropriate for modeling the data.
- When we examined the boxplot of claim amounts, we observed a significant presence of outliers, indicating considerable skewness in the data.

RESULT AND DISSCUSION

OBJECTIVE 1

For case 1 :

Identify which risk factors are most important in predict likelihood of claim to help actuaries.

#chi-squared test

H0: Region and insurance claim are independent.

H1: Region and insurance claim are dependent.

```
        Pearson's Chi-squared test

data:  data$region and data$insuranceclaim
X-squared = 21.679, df = 3, p-value = 7.606e-05
```

Interpretation:-

The low p-value (0.00007606) confirms a significant link between region and claim amounts, suggesting insurers should consider regional factors in pricing and risk management.

- The table below presents the results of a Generalized Linear Model (GLM) fitted to predict insurance claim amounts.

Variable	Estimate	Std.Error	Z value	Pr(>|t|)
(Intercept)	-7.387e+00	5.792e-01	-12.755	<2e-16***
age	2.677e-02	7.257e-03	3.690	0.000225
Sex1	-1.574e-02	1.584e-01	-0.099	0.920836
bmi	2.586e-06	1.821e-02	14.204	<2e-16***
children	-1.424e+00	9.429e-02	-15.106	<2e-16***
smoker	4.048e+00	4.097e-01	9.644	<2e-16***
region	-9.372e-02	7.210e-002	-1.300	0.193642
charges	5.800e-06	1.553e-05	0.374	0.708774

Significance Codes: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.', 0.1 ' ' 1

Null Deviance: 1815.82 on 1337 degrees of freedom.

Residual Deviance: 997.5 on 1330 degrees of freedom.

AIC: 1013.1

Number of Fisher Scoring Iterations: 6

- Below is the interpretation of the odds ratio estimates:

Effects	Odds ratio estimates
Age	1.027136e+00
Sex	9.843833e-01
Bmi	1.295146e+00
Children	2.406667e-01
Smoker	5.728467e+01
Region	9.105341e-01
Charges	1.000006e+00

Interpretation: -

- Smoking is the most important risk factor in predicting claim likelihood. Actuaries should heavily weigh smoking status when assessing risk and setting insurance premiums.
- BMI is also an important factor, indicating a link between obesity and claim probability.
- Age has a moderate impact, meaning older individuals are slightly more likely to file claims.
- Children and region have minimal or inverse effects, suggesting demographic factors might not be as strong predictors.
- Sex and charges do not play a major role in claim likelihood, based on the model.

Overall objective1 conclusion:

Smoking, BMI, and age are the most important risk factors for predicting insurance claims. **Smokers are 57 times more likely** to file a claim, making it the strongest predictor. **Higher BMI significantly increases risk**, and **older individuals have a slightly higher claim likelihood**. Other factors like sex, region, and prior charges have minimal impact. **Actuaries should prioritize smoking, BMI, and age in risk assessment and premium pricing.**

OBJECTIVE 2

Determine a base premium and adjust it according to an individual's risk category (Low, Medium, High).

The base premium is calculated before adjusting for risk factors.

$$Base\ premium = \frac{mean(charges)}{12}$$

- Mean(charges) = 13270.42
- Base premium= 1105.869

- Criteria Consider for Risk Categories from Muthoot Finance

Risk Category	Risk Factor	Criteria Considered
Low Risk	1.0	• Non-smoker • BMI within a healthy range • Low medical charges
Medium Risk	1.5	• Slightly higher BMI • Moderate medical charges • Younger smoker
High Risk	2.5	• Smoker • High BMI • Older age

The adjusted premium calculated by

$$adjusted\ premium = \frac{base\ premium}{risk\ factor}$$

Here is the adjusted premium calculated with the help of risk category or risk factor

	age	sex	bmi	children	smoker	region	charges	insuranceclaim	risk_category	risk_category_numeric	risk_factor	adjusted_premium
1	19	0	27.900	0	1	3	16884.924	1	high risk	2	2.5	2764.671
2	18	1	33.770	1	0	2	1725.552	1	medium risk	1	1.5	1658.803
3	28	1	33.000	3	0	2	4449.462	0	low risk	0	1.0	1105.869
4	33	1	22.705	0	0	1	21984.471	0	medium risk	1	1.5	1658.803
5	32	1	28.880	0	0	1	3866.855	1	medium risk	1	1.5	1658.803
6	31	0	25.740	0	0	2	3756.622	0	medium risk	1	1.5	1658.803
7	46	0	33.440	1	0	2	8240.590	1	medium risk	1	1.5	1658.803
8	37	0	27.740	3	0	1	7281.506	0	low risk	0	1.0	1105.869
9	37	1	29.830	2	0	0	6406.411	0	low risk	0	1.0	1105.869
10	60	0	25.840	0	0	1	28923.137	0	medium risk	1	1.5	1658.803
11	25	1	26.220	0	0	0	2721.321	1	medium risk	1	1.5	1658.803
12	62	0	26.290	0	1	2	27808.725	1	high risk	2	2.5	2764.671
13	23	1	34.400	0	0	3	1826.843	1	medium risk	1	1.5	1658.803
14	56	0	39.820	0	0	2	11090.718	1	high risk	2	2.5	2764.671
15	27	1	42.130	0	1	2	39611.758	1	high risk	2	2.5	2764.671

Interpretation: -

- Low Risk individuals are generally healthier they are less likely to file insurance claims. These individuals receive lower premium
- Medium Risk individuals have moderate health concerns, possibly due to slightly higher BMI. They are more likely to incur medical costs than low-risk individuals, so their insurance premiums are adjusted accordingly.
- High Risk individuals represent older age, Smoker, moderate BMI or high BMI and due to older age increased risk of significant medical expenses. This risk-based approach helps insurance companies maintain a balanced and profitable model while ensuring that policyholders receive appropriate premium rates.

Overall objective2 conclusion: -

By categorizing individuals into Low, Medium, and High-Risk segments based on factors such as age, smoking status, and BMI, insurers can more accurately align premiums with the likelihood of incurring medical costs. Low Risk individuals benefit from lower premiums due to healthier profiles and reduced claims. Medium Risk individuals, with moderate health concerns, pay adjusted premiums that reflect their slightly increased likelihood of medical expenses. High Risk individuals, often older or smokers with higher BMIs, face higher premiums because of their elevated probability of significant healthcare costs. This tiered, risk-based strategy enables insurers to maintain a balanced, profitable model while ensuring that premiums fairly match each policyholder's risk level.

OBJECTIVE 3

For case 2

To Check the Association between the variables by welch two sample t test

- **bmi by smoker**

Welch Two Sample t-test

data: bmi by smoker
t = -3.7055, df = 4132.7, p-value = 0.0002137
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
-0.7472778 -0.2301408
sample estimates:
mean in group 0 mean in group 1
30.19047 30.67918

Interpretation: -

The t-test shows that smokers have a statistically significantly higher BMI (about 0.23 to 0.75 units more) than non-smokers. The p-value is 0.0002137, which is well below the common significance level of 0.05, indicating that the difference is statistically significant.

- **claim by smoker**

Welch Two Sample t-test

data: claim by smoker
t = -97.405, df = 3078.2, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
-23696.8 -22761.6
sample estimates:
mean in group 0 mean in group 1
8814.197 32043.395

Interpretation: -

The t-test indicates a highly significant difference in claim amounts between non-smokers and smokers. Non-smokers average about 8,814 while smokers average around 32,043. The difference (approximately 23,000) is statistically significant, confirming that smokers tend to have much higher claims.

- **claim by diabetes**

Welch Two Sample t-test

data: claim by diabetes
t = -18.996, df = 6742, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
-4259.133 -3462.293
sample estimates:
mean in group 0 mean in group 1
10393.07 14253.78

Interpretation: -

The t-test shows that diabetics have significantly higher claim amounts than non-diabetics, with an average difference of roughly 3,500 to 4,260 units, and the result is statistically significant**.** The extremely small p-value (< 2.2e-16) strongly indicates that the observed difference is not due to random chance.

OBJECTIVE 4

To handle under dispersion in a skewed response variable using a quasi-likelihood approach and ensure reliable modelling with a constant coefficient of variation.

- The table below presents the results of a Generalized Linear Model (GLM) fitted to predict insurance claim amounts.

Variable	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	1.668e-04	2.740e-6	60.873	$< 2e{-}16$***
Age	-1.068e-06	2.569e-08	-41.579	$< 2e{-}16$***
Sex	1.553e-06	6.507e-07	2.387	0.0170*
bmi	-7.335e-07	5.689e-08	-12.893	$< 2e{-}16$***
Hereditary DiseasesArthritis	3.966e-06	1.606e-06	2.469	0.0136
Hereditary DiseasesCancer	1.720e-05	1.915e-06	8.978	$< 2e{-}16$***
Hereditary DiseasesDiabetes	1.0336e-05	1.799e-06	5.758	8.72e-09***
Hereditary DiseasesEpilepsy	2.400e-	2.178e-06	11.019	$< 2e{-}16$***
Hereditary DiseasesEyeDisease	1.36705	1.828e-06	7.479	7.96e-14***
Hereditary DiseasesHeartDisease	6.256e-05	1.801e-06	0.347	0.7283
Hereditary DiseasesHigh BP	1.203e-07	2.196e-06	5.480	4.33e-08***
Hereditary DiseasesNoDisease	1.787e-05	1.405e-06	12.712	$< 2e{-}16$***
Hereditary DiseasesObesity	1.600e-05	1.873e-06	8.542	$< 2e{-}16$***
No of Dependents	-2.580e-05	2.653e-07	-9.727	$< 2e{-}16$***
Smoker	-7.341e-06	8.417e-07	-87.221	$< 2e{-}16$***
Diabetes	-5.842e-05	1.063e-06	-5.498	3.19e-08***
bloodpressure	4.012e-06	1.785e-08	2.247	*0.0246**

Significance Codes: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.', 0.1 ' ' 1

Null Deviance: 10517.1on 13647 degrees of freedom.

Residual Deviance: 4953.8 on 13631 degrees of freedom.

AIC: 274669

Number of Fisher Scoring Iterations: 6

Interpretation:

- Intercept) = 1.668e-04

  The expected claim amount ,when all predictors are 0

- Age = -1.068e-06

  Older individuals tend to have lower claim amounts.

- Sex = 1.553e-06

  Gender has a small but significant effect on claims.

- BMI = 0.009154:

  Higher BMI is associated with slightly lower claim amounts.

- Hereditary_diseasesArthritis =3.966e-06

People with a hereditary disease of arthritis have an expected claim amount is higher

- Hereditary_diseasesCancer = 1.720e-05

  Significantly increases claim amount.

- Hereditary_diseasesdiabetes =1.036e-05

  Increases claim amount significantly

- Hereditary_diseaseseilepsy=2.400e-05

  Strongly increases claim amount.

- Hereditary_diseasesobesity=1.367e-05

Increases claim amount.

- Hereditary_diseaseshigh bp=-1.203e-05

Increases claim amount.

- Hereditary_diseaseseye=1.367e-05

Increases claim amount

- Hereditary_diseasesheart=6.256e-07

  No clear effect on claim amount.

- No of Dependents = -2.580e-06

  More dependents lead to lower claims

- Smoker = -7.341e-05

The negative coefficient suggests that being a smoker is associated with lower claim amounts.

- Diabetes = -5.842e-06

  having diabetes reduces the expected claim amount.

- Blood Pressure=4.012e-08

  Small but significant increase in claim amount.

Interpretation:

Demographics: Older individuals and those with more dependents tend to have **lower** claim amounts. Gender has a minor effect in given data

Health & Lifestyle: Higher BMI, smoking, and diabetes unexpectedly correlate with lower claim amounts, suggesting possible behavioral or policy influences from this data

Hereditary Diseases: Conditions like cancer, diabetes, epilepsy, obesity, high blood pressure, and eye disease significantly increase claim amounts, with epilepsy having the strongest impact. Heart **disease** shows no clear effect.

Medical Indicators: Blood pressure has a small but significant positive effect on claims.

dispersion parameter: 0.2640158

Here the dispersion parameter < 1, it indicates under dispersion.

- **Use a Quasi-Likelihood Model, for an adjustable dispersion parameter**

Variable	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	8.0049975	0.0279417	286.489	< 2e−16***
Age	0.0140914	0.0002367	59.526	< 2e−16***
Sexmale	-0.0079810	0.0062862	-1.270	0.204245
Bmi	0.0206554	0.0005252	39.328	< 2e−16***
Smoker	1.1736640	0.0076528	153.364	< 2e−16***
diabetes	0.1849548	0.0098578	18.762	< 2e−16***
bloodpressure	0.0005524	0.0001678	3.291	0.000999*
Hereditary DiseasesArthritis	0.0634301	0.0241320	2.628	0.008587
Hereditary DiseasesCancer	-0.1939291	0.0209720	-9.247	< 2e−16***
Hereditary DiseasesDiabetes	-0.1251684	0.0192716	-6.490	8.89e-11
Hereditary DiseasesEpilepsy	-0.2902064	0.0252561	-11.491	< 2e−16***
Hereditary DiseasesEyeDisease	-0.1218764	0.0192716	-6.324	*2.63e-10*
Hereditary DiseasesHeartDisease	0.0320741	0.0202859	1.581	0.113878
Hereditary DiseasesHigh BP	-0.1318142	0.0227033	-5.806	6.54e-09
Hereditary DiseasesNoDisease	-0.3184276	0.0147883	-21.523	< 2e−16***
Hereditary DiseasesObesity	-0.1409317	0.0201607	-6.990	2.87e-12***
No of dependents	0.0218375	0.0026105	8.365	< 2e−16***

Significance Codes: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.', 0.1 ' ' 1

Null Deviance: 1.9915e+12 on 13647 degrees of freedom.

Residual Deviance: 4.7898e+11 on 13631 degrees of freedom.

AIC: NA

Number of Fisher Scoring Iterations: 7

dispersion parameter: 35139021

#Here the dispersion parameter >1, it indicates overdispersion

- **The fitted model is using inverse link is**

$claim^{-1}$=1.668×10−4+(−1.068×10−6×age)+(1.553×10−7×sex)+(−7.335×10−7 ×bmi)+(3.966×10−6×hereditary_diseases(Arthritis))+(1.720×10−5×hereditary_ diseases(Cancer))+(1.036×10−5×hereditary_diseases(Diabetes)) +(2.400×10−5×hereditary_diseases(Epilepsy))+(1.367×10−5×hereditary_diseas es(Eye Disease)) +(6.256×10−7×hereditary_diseases(Heart Disease))+(1.203×10−5×hereditary_d iseases(High BP)) +(1.787×10−5×hereditary_diseases(No Disease))+(1.600×10−5×hereditary_dis eases(Obesity)) +(−2.580×10−7×no_of_dependents)+(−7.341×10−7×smoker) +(−5.842×10−6×diabetes)+(4.011×10−8×bloodpressure)

Interpretation:

The initial dispersion parameter (**0.264**) indicated **under dispersion**, meaning claim amounts varied less than expected. By applying a **quasi-likelihood model**, the dispersion parameter increased significantly (**35,139,021**), resulting in **overdispersion**. This transformation allows the model to better capture real-world variability, improving prediction accuracy and risk assessment.

OBJECTIVE 5

Recognizing patterns that indicate suspicious or fraudulent claims

- Train Random Forest model

Call:

randomForest(formula = is_fraud ~ claim + age + bmi + smoker + bloodpressure + diabetes, data = train, ntree = 100)

Type of random forest: regression

Number of trees: 100

No. of variables tried at each split: 2

Mean of squared residuals: 6.334481e-05

% Var explained: 99.87

OOB estimate of error rate: 0%

Confusion matrix:

	0	1	class. Error
0	9070	0	0
1	0	483	0

Interpretation: -

(MSE = 6.334481e-05) This value represents the average squared difference between the predicted and actual fraud values.A very small MSE suggests that the model makes accurate predictions with minimal error.

Variance Explained (99.87%):The model explains 99.87% of the variation in the fraud detection outcome.This indicates a highly effective model with strong predictive power.

OOB (Out-of-Bag) error rate = 0%, meaning the model made no mistakes when predicting fraud.

Confusion Matrix Interpretation:

- 9,070 genuine claims were correctly labeled as not fraud (0).
- 483 fraudulent claims were correctly identified as fraud (1).
- No misclassifications

CONCLUSION

Smokers are 57 times more likely to file insurance claims, and other factors like higher BMI and older age also increase risk. Using this data, insurers can set fair premiums based on actual risk levels. To ensure fairness, a three-tier risk system (Low/Medium/High) was introduced, allowing healthier individuals to pay lower premiums while high-risk groups, such as smokers and older adults, pay more. Fraud detection was also enhanced using a Random Forest model, which achieved 99.87% accuracy and successfully identified all 483 fraud cases. Statistical tests revealed that smokers have significantly higher claims (₹32,043 on average compared to ₹8,814 for non-smokers) and that diabetics and individuals with hereditary diseases, such as cancer, tend to file costlier claims. To further improve accuracy, advanced models like Gamma GLM were used to correct skewed data and ensure reliable predictions. These data-driven strategies help insurers predict risks more accurately, set fair prices so healthier individuals don't overpay, detect fraud to keep overall costs down, and build trust through transparent, evidence-based decision-making.

APPENDIX

For case 1:

```
data=read.csv(file.choose(),header=TRUE)

data

head(data)

summary(data)

library(ids)

sum(is.na(data))

data_clean=na.omit(data)

data_clean

chisq_test=chisq.test(data$region,data$insuranceclaim)

chisq_test

glm_model=glm(insuranceclaim~age+sex+bmi+children+smoker+region+char
ges,data=data,family=binomial(link="logit"))

summary(glm_model)

mean(predicted==data$insuranceclaim)

predicted_risk=predict(glm_model2,type = "response")

predicted_risk

mean(data$charges)

base_premium=mean(data$charges)/12

base_premium

bins=c(0.0004281,0.2570631,0.9359493,0.9999679)

labels=c("low risk","medium risk","high risk")

labels

data$risk_category=cut(predicted_risk,breaks=bins,labels=labels,right=FALSE)

data$risk_category

data$risk_factor=ifelse(data$risk_category == "low risk", 1.0,
```

```
 ifelse(data$risk_category == "medium risk", 1.5, 2.5))
data$risk_factor
data$adjusted_premium = base_premium * data$risk_factor
data$adjusted_premium
print(data)
```

For case 2:

```
data1=read.csv(file.choose(),header=T)
data1
View(data1)
dim(data1)
str(data1)
summary(data1)
boxplot(data1$claim)
hist(data1$claim)
t.test(bmi ~ smoker, data = data1)
t.test(claim ~ smoker, data = data1)
t.test(claim ~ diabetes, data = data1)
y=data1$claim;y
cv_y=sd(y)/mean(y);cv_y
alpha=1/cv_y^2
alpha
s=mean(y)/alpha                #scale parameter of gamma distribution
s
ks.test(y,alpha,s,exact=NULL,'gamma')
bloodpressure=data1$bloodpressure
bloodpressure
fitted_model=glm(claim~age+sex+bmi+hereditary_diseases+no_of_dependents
+smoker+diabetes+bloodpressure,data=data1,family=Gamma(link="inverse"))
```

```
s=summary(fitted_model)
s
df_residual=df.residual(fitted_model)
df_residual
dispersion_parameter=residual_deviance / df_residual
dispersion_parameter
beta=s$coefficients
beta=as.vector(beta)
beta
x0=c(1,1,1,1,1,1,1,1,1,1,1,1,1)
x0
eta=beta*x0
eta
mu=1/as.numeric(beta)
mu
influence.measures(fitted_model)
fitted_model1=glm(claim~age+sex+bmi+hereditary_diseases+no_of_dependent
s+smoker +diabetes+bloodpressure,data=data1,family=Gamma(link="log"))
summary(fitted_model1)
glm_quasi=glm(claim ~ age + sex + bmi + smoker + diabetes +
bloodpressure+hereditary_diseases+no_of_dependents, family = quasi(link =
"log"), data = data1)
summary(glm_quasi)
residual_dgeviance=4.7898e+11
df_residual=df.residual(glm_quasi)
df_residual
dispersion_parameter=residual_deviance / df_residual
dispersion_parameter
```

```
library(randomForest)
library(dplyr)

data=read.csv(file.choose(),header=T)
data
data$claim = as.numeric(data$claim)
threshold =quantile(data$claim, 0.95, na.rm = TRUE)
# 1 = Fraud, 0 = Genuine
data$is_fraud =ifelse(data$claim > threshold, 1, 0)
data$is_fraud
data$smoker =as.factor(data$smoker)
data$diabetes = as.factor(data$diabetes)
data$is_fraud = as.factor(data$is_fraud)
set.seed(123)
trainIndex=sample(1:nrow(data), 0.7 * nrow(data))
trainIndex
train=data[trainIndex, ]
train
test=data[-trainIndex, ]
test
mean(as.numeric(as.character(test$is_fraud))))^2)
sst=sum((as.numeric(as.character(test$is_fraud)) -
sse  sum((test$predicted_fraud - as.numeric(as.character(test$is_fraud)))^2)
r_squared = 1 - (sse / sst)
r_squared
rf_model= randomForest(is_fraud ~ claim + age + bmi + smoker +
bloodpressure + diabetes, data = train, ntree = 100)
rf_model
```

```
predictions=predict(rf_model, test)
predictions
conf_matrix = table(Predicted = predictions, Actual = test$is_fraud)
print(conf_matrix)
```

REFERANCE

Agresti, A. (2002). Categorical Data Analysis, ED.II, Wiley InterScience

McCullagh, P. And Nelder,J.A. (1983). Generalized Linear Models- Monographs on Statistics and Applied Probability, Chapman and Hall

Myers, R.H, Montegomery, D.C., Vinning, G.G and Robinson, T.J.(2010). Generalized Linear Models with Applications in Engineering and the Sciences, Ed.II , Wiley Series in Probability and Statistics, A John Wiley & Sons.