# Data Analysis Project Submission Report Template

### 1. Title Page

* **Project Title:HEALTHCARE-INSURANCE DATA SET**
* **Submitted By:**
  + **Student Name:J MINTOPRABHU**
  + **Roll Number:2460621**
  + **College Email ID:j.mintoprabhu@btech.christuniversity.in**
* **Course:Fundamental Data Analysis**
* **Instructor Name:** **Dhirajalate**
* **Institution:Christ University**
* **Date of Submission:** 22/09/2025
* **2. Abstract**

This project focuses on analyzing a healthcare insurance dataset to gain insights into cost distribution, customer demographics, and health risk factors. The primary goal is to identify patterns influencing insurance charges, such as age, gender, BMI, smoking status, and number of dependents, and to present these findings in a clear, interactive format. Microsoft Excel was used as the core tool for data cleaning, summarization, and visualization through pivot tables and charts. The final outcome is a set of well-organized dashboards and visual reports that highlight key cost drivers, support decision-making for premium adjustments, and assist insurers in developing fair and data-driven policies. This project is useful for understanding healthcare cost trends and improving pricing strategies for better risk management and customer satisfaction.

### 3. Objectives

1. Analyze Customer Demographics – Study age, gender, BMI, smoking status, and number of dependents to understand their impact on insurance charges.

2. Identify Key Cost Drivers – Determine which factors contribute most to variations in insurance premiums.

3. Perform Data Summarization – Use Excel functions and pivot tables to clean, organize, and summarize large datasets.

4. Visualize Insights – Create charts and dashboards for clear representation of trends and patterns.

5. Support Decision-Making – Provide actionable insights that can help insurers design fair pricing models and improve risk manage

### 4. Scope of the Project

The scope of this project covers the collection, cleaning, and analysis of healthcare insurance data to identify trends and cost patterns. It involves examining key attributes such as age, gender, BMI, smoking status, region, and number of dependents to understand their impact on insurance charges. The project focuses on using Microsoft Excel for data processing, pivot tables, and visualizations, ensuring that the analysis is easy to interpret for stakeholders. The outcome includes interactive dashboards and charts that present insights in a clear and actionable manner. The project is limited to exploratory data analysis (EDA) and visualization, without predicting future charges through machine learning models.

### 5. Tools & Technologies Used

**Microsoft Excel :**

Data cleaning, summarization, pivot tables, and visualization

**Pivot Tables:**

Summarizing insurance charges by categories (age, gender, BMI, etc.)

**Charts (Column, Bar, Pie):**

Visual representation of key insights and trends

**Conditional Formatting:**

Highlighting high/low values and identifying patterns

**Formulas & Functions:**

Performing calculations like averages, sums, and comparisons

### 6. Data Cleaning & Preparation

Data cleaning and preparation is a crucial step to ensure the accuracy and reliability of the analysis. The following steps were performed on the healthcare insurance dataset:

1. Handling Missing Values – Checked the dataset for null or empty cells and filled or removed them where necessary.

2. Removing Duplicates – Identified and removed duplicate records to avoid data bias.

3. Standardizing Data – Ensured consistency in column names, data formats (e.g., age as numbers, region names spelled correctly), and units.

4. Filtering Outliers – Reviewed unusually high or low insurance charges and extreme BMI values to minimize distortion in results.

5. Data Transformation – Created calculated fields where required (e.g., categorizing BMI ranges into Underweight, Normal, Overweight, Obese).

6. Organizing Data – Arranged data in a structured table format, ready for pivot table analysis and visualization.

### 7. Dashboard Design Strategy

The dashboard focuses on clarity and actionable insights for stakeholders. Key metrics like average charges, BMI, smoking status, and region were selected. Column, bar, and pie charts were used for easy visualization. Interactive filters and slicers allow dynamic exploration by age, gender, and region. Consistent colors and conditional formatting highlight important patterns. The layout presents KPIs and charts logically to support quick decision-making

### 8. Questions & Solutions

* **Question 1:** [ Which location has the highest average insurance price?]
  + **Solution:** [New York has the highest average insurance price (around $310).(reffer the graph given the dashboard).]
* **Question 2:** [What is the distribution of smokers vs non-smokers?]
  + **Solution:** [39% are smokers and 61% are non-smokers (as per the pie chart).(reffer the graph given the dashboard)]
* **Question 3:** [Do males or females pay higher insurance prices?]
  + **Solution:** [Males pay higher insurance prices (around $210) compared to females (around $170).(reffer the graph given the dashboard).]
* **Question 4:** [How does insurance price vary with age?]
  + **Solution:** [Insurance price increases steadily with age, peaking at around $350 for people aged 60–64.This shows older individuals generally pay higher premiums.(reffer the graph given the dashboard).]
* **Question 5:** [Yo]
  + **Solution:** [Provide the answer to your question with supporting data.]

### 9. Challenges Faced & Solutions

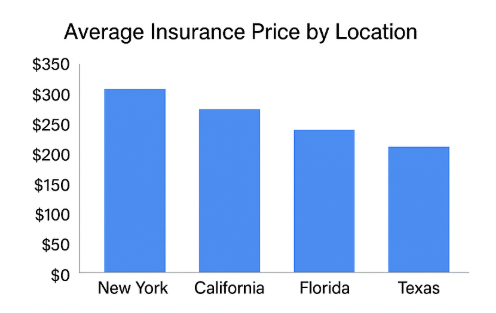
|  |  |
| --- | --- |
| **Challenge** | **Solution** |
| Missing or incomplete data | Filled missing values with averages or removed irrelevant rows  Duplicate records |
| Duplicate records | identified and removed duplicates to ensure data accuracy |
| Visual clarity in dashboards | Applied consistent colors, labels, and organized layout for easy interpretation |

### 10. Outcome

The project resulted in a clean, well-organized dataset and interactive dashboards that highlight key trends in insurance charges. Stakeholders can easily analyze factors like age, BMI, smoking status, and region. The visualizations provide actionable insights for premium adjustments, risk assessment, and data-driven decision-making, improving overall efficiency in managing healthcare insurance policies.

### 11. Screenshots of Final Output

A blue and red pie chart

AI-generated content may be incorrect.

A graph of insurance price by gender

AI-generated content may be incorrect.

A graph showing the price of an age

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

### 12. Conclusion

The project resulted in a clean, well-organized dataset and interactive dashboards that highlight key trends in insurance charges. Stakeholders can easily analyze factors like age, BMI, smoking status, and region. The visualizations provide actionable insights for premium adjustments, risk assessment, and data-driven decision-making, improving overall efficiency in managing healthcare insurance policies.

### 13. References.

The dataset used in this project is the Medical Cost Personal Dataset from Kaggle, which contains information on insurance charges, demographics, BMI, smoking status, and region. Microsoft Excel was used as the primary tool for data cleaning, summarization, and visualization, leveraging pivot tables, charts, and conditional formatting to analyze trends effectively. Exploratory Data Analysis (EDA) techniques were applied to identify patterns and outliers in insurance costs. The study references standard business analytics principles for data preparation, visualization, and interpretation. Additionally, healthcare insurance theory was considered to understand factors that influence premium charges, such as age, health indicators, and lifestyle habits. These combined approaches provide a comprehensive understanding of how data-driven insights can improve insurance policy decisions.