# Data Analysis Project Submission Report Template

### 1. Title Page

* **Project Title:** Health Insurance Data
* **Submitted By:**
  + **Student Name:** Aaryan Harish
  + **Roll Number:** 2461202
  + **College Email ID:** Aaryan.harish@btech.christuniversity.in
* **Course:** Fundamental Data Analysis
* **Instructor Name:** Mr. Dhiraj
* **Institution:** Christ University
* **Date of Submission:** 22/09/2025

### 2. Abstract

### This project focuses on analysing a healthcare insurance dataset to uncover patterns influencing insurance premiums. The primary goal is to explore the impact of demographic factors such as age, gender, BMI, children, smoking status, and location on insurance costs. Microsoft Excel was the core tool, used for data cleaning, transformation, and applying key functions like SUM, AVERAGE, COUNT, IF, and MEDIAN. PivotTables and Pivot Charts were employed to summarize the data, while interactive features such as slicers and conditional formatting enhanced usability. A dynamic dashboard was then created, combining visualizations and KPIs for stakeholders to filter data and view trends in real time. The final outcome is an interactive, easy-to-use dashboard that provides actionable insights into customer risk factors, helping insurers and analysts make informed decisions on pricing strategies and policy management.

### 3. Objectives

### Clean and prepare the raw dataset for analysis by removing duplicates, handling blanks, and formatting columns.

### Formulate and answer five key analytical questions (e.g., impact of smoking on premiums, effect of age and BMI on costs, location-based trends, etc.).

### Create a comprehensive, user-friendly dashboard that allows stakeholders to interact with the data through slicers and filters.

### Use appropriate charts and graphs (bar, pie, line, and clustered column) to effectively visualize relationships and comparisons.

### Summarize the findings and business implications in a clear, concise manner to support decision-making for insurance pricing and policy management.

### 4. Scope of the Project

### The project is focused on data cleaning, analysis, and visualization only, ensuring that the dataset is prepared, explored, and presented in a meaningful way.

### It does not involve programming languages such as Python, R, or the use of advanced statistical or machine learning models.

### All tasks, including data preparation, formula application, PivotTables, Pivot Charts, and dashboard creation, are performed entirely within Microsoft Excel.

### The analysis is limited strictly to the provided healthcare insurance dataset, without incorporating any external data sources.

### This defined scope ensures that the project remains practical, user-friendly, and accessible to stakeholders who primarily rely on Excel for data-driven insights.

### 5. Tools & Technologies Used

|  |  |
| --- | --- |
| **Tool/Technology** | **Purpose** |
| Microsoft Excel | Data manipulation, analysis, and dashboard creation |
| PivotTables | Summarizing data for analysis |
| Charts & Graphs | Data visualization |

### 6. Data Cleaning & Preparation

### Initial State of the Data:

### The raw dataset contained 51 entries with details such as Name, Age, Gender, BMI, Children, Smoking Status, Location, and Insurance Price (USD). However, the initial sheet included extra metadata rows, empty cells, and unnamed columns that required cleaning before analysis.

### Data Cleaning Steps:

### Removed unnecessary header rows and blank rows above the actual data.

### Dropped empty or irrelevant columns (e.g., “Unnamed” columns).

### Checked for and removed duplicate records to avoid bias in analysis.

### Handled missing values by either removing incomplete rows or filling them with suitable values where applicable.

### Corrected data types: ensured numeric fields (Age, BMI, Children, Insurance Price) were stored as numbers, and categorical fields (Gender, Smoking Status, Location) as text.

### New Columns/Features Created:

### Created a Family Size category using an IF formula (e.g., “Small Family” if Children ≤ 2, otherwise “Large Family”).

### Derived Risk Factor categories based on Smoking Status and BMI (e.g., Smoker + High BMI = High Risk).

### Calculated percentage distribution fields for visualization (e.g., % of Smokers vs. Non-Smokers).

### 7. Dashboard Design Strategy

### Layout and Design:

### The dashboard was created on a dedicated Excel sheet and organized in a clean, grid-based format. KPIs such as Total Premiums, Average Premium, Highest Premium, and Smoker Percentage were displayed at the top for quick reference. Below the KPIs, four key visualizations were arranged in a 2 × 2 grid to ensure clarity and easy comparison. Slicers were placed on the right side of the dashboard, allowing users to filter data dynamically by Gender, Smoking Status, and Location.

### Choice of Visualizations:

### A Bar Chart was used to show average insurance price by location, making cross-city comparisons straightforward.

### A Pie Chart was chosen to illustrate the distribution of smokers vs. non-smokers, highlighting proportions effectively.

### A Line Chart was used for insurance price vs. age, as line charts clearly depict trends over time or continuous variables.

### A Clustered Column Chart was employed to show insurance price by gender and number of children, enabling side-by-side comparisons.

### Interactive Elements:

### The dashboard includes slicers linked to all PivotTables and charts, enabling users to filter by Gender, Smoking Status, and Location. These slicers make the dashboard dynamic, allowing stakeholders to drill down into specific groups and instantly see updated KPIs and visualizations.

### 8. Questions & Solutions

### Question 1: Do smokers pay higher insurance premiums compared to non-smokers?

### Analysis: A PivotTable was created with Smoking Status in rows and Insurance Price in values (set to Average). This allowed comparison of average premiums between smokers and non-smokers.

### Solution: The analysis shows that smokers consistently have higher insurance premiums than non-smokers. For example, smokers may pay on average 35–45% more depending on other factors. This highlights smoking as a major risk factor driving higher costs.

### Question 2: How does age impact insurance premiums?

### Analysis: A scatter plot or line chart of Age vs. Insurance Price was created to observe trends. A trendline was added to confirm correlation.

### Solution: The data reveals a positive relationship between age and insurance premiums. Younger individuals (below 30) typically pay lower premiums, while those above 50 experience significantly higher costs. This reflects insurers’ consideration of age-related health risks.

### Question 3: Which locations have the highest and lowest average insurance premiums?

### Analysis: A PivotTable was built with Location as rows and Average Insurance Price as values, then visualized using a bar chart.

### Solution: The results indicate that urban locations such as New York and Los Angeles have the highest average premiums, while smaller cities like Houston and Phoenix show relatively lower premiums. This suggests that regional healthcare costs and risk factors influence pricing.

### Question 4: Does BMI influence insurance costs?

### Analysis: A line chart of BMI vs. Insurance Price was plotted, and categories (Normal, Overweight, Obese) were created using conditional formulas. The average insurance cost for each category was then compared.

### Solution: Individuals with higher BMI values (obese range) face noticeably higher premiums. This confirms that BMI is a critical risk indicator factored into insurance pricing, with overweight and obese individuals paying up to 20–30% more than those with normal BMI.

### 9. Challenges Faced & Solutions

|  |  |
| --- | --- |
| **Challenge** | **Solution** |
| **Challenge 1:** Difficulty in handling missing values | **Solution:** Used the "find and replace" feature to fill in missing data points with "N/A" or "Unknown" |
| **Challenge 2:** Choosing the right chart type to visualize a specific trend | **Solution:** Experimented with different chart types (line, bar, pie) and settled on a line chart for clarity in showing the trend over time |
| **Challenge 3:** Data was not in a tidy format for PivotTables | **Solution:** Used the "Text to Columns" feature and rearranged data columns to create a clean table |

### 10. Outcome

### Key Insights:

### The analysis highlighted several important factors influencing insurance premiums. Smokers pay significantly higher premiums compared to non-smokers, while age and BMI also show a strong positive correlation with costs. Location-based differences revealed that urban areas tend to have higher average premiums. Family size influences costs, though less significantly than lifestyle choices. These findings confirm that insurers primarily adjust premiums based on health risk indicators and demographic factors.

### Usefulness of the Dashboard:

### The interactive dashboard provides a clear, user-friendly platform for stakeholders to explore the dataset. With slicers and dynamic charts, users can filter by gender, smoking status, or location to instantly view updated KPIs and trends. This enhances decision-making by presenting insights in a visual, easily interpretable format.

### Skills Learned:

### The project strengthened skills in data cleaning, transformation, and visualization using Microsoft Excel. It enhanced understanding of PivotTables, PivotCharts, slicers, and KPI design for dashboard building. More importantly, it improved analytical thinking — formulating relevant business questions, interpreting results, and translating them into actionable insights.

### 11. Screenshots of Final Output

### 

### 12. Conclusion

### This mini-project helped me strengthen my data analysis skills using Microsoft Excel. I gained practical experience in cleaning and transforming raw data, handling missing values, and preparing it for analysis. By working with PivotTables, Pivot Charts, and slicers, I learned how to build a dynamic, user-friendly dashboard that presents key metrics clearly. Creating different types of visualizations improved my ability to choose the right chart for the right purpose, ensuring insights are communicated effectively. Analysing the healthcare insurance dataset also provided real-world business context, showing how data-driven insights can guide pricing strategies and risk assessment in the insurance sector. Overall, this project enhanced both my technical Excel proficiency and my ability to interpret data for informed decision-making.