Exercise Lab 6

Mina Moeini
Mina_moeini@sfu.ca

October 18, 2024

Instructions

Using the provided healthcare dataset, analyze the data to answer the following questions. Use appropriate statistical methods to compute the requested performance measures. Present your results with necessary calculations and interpretations.

Questions

1. Average Billing for each medical condition

- What is the mean and standard deviation of the billing amount for each medical condition?
- Based on the horizontal bar chart:
  - Which medical condition has the highest number of observations?
  - Which condition has the lowest number of observations?
- Using the box plot generated the billing amount for each group to show the standard deviation and sd.

2. Exploring the Relationship between Medical Condition and Blood Type

- Bar Chart: For each medical condition, create a grouped bar chart that compares the blood types and their corresponding average billing amounts.
- What patterns do you observe in the grouped bar charts

3. Average Billing Performance

- What is the average billing amount for each hospital?
- How does the average billing amount compare between urgent and elective admissions?

Hint: Group the data by the `Hospital` and `Admission Type` columns, and calculate the mean of the `Billing Amount` column.

4. Admission Outcome Analysis

- What is the average length of stay for each type of admission (urgent, elective, or emergency)?

Hint: Calculate the difference between `Discharge Date` and `Date of Admission`, and find the average for each `Admission Type`.

5. Insurance Billing Performance

- Which insurance provider has the highest average billing amount?

Hint: Group the data by `Insurance Provider` and calculate the mean of the `Billing Amount` column.

6. Gender-Based Analysis

- What is the average billing amount by gender?
- Are there any significant differences in billing between males and females?

Hint: Group the data by `Gender` and compare the average `Billing Amount`.

7. Age Group Analysis

- Define the following age groups: Young (0–40), Middle-aged (41–65), and Senior (66+).
- Calculate the total billing amount for each age group.

Hint: Create age bins and sum the `Billing Amount` for each bin.

More expansion over question1.

1. Data Import and Structure Inspection

(a) What SAS procedure is used to import the healthcare dataset?

(b) After importing, how can you check the structure and column names of the dataset? Provide the procedure and explain its output.

2. Average Billing by Hospital

(a) Which procedure is used to calculate the average billing amount for each hospital?

(b) How are the top 10 most expensive hospitals selected from the results?

3. Visualization: Top 10 Hospitals

(a) What type of plot is used to visualize the top 10 hospitals by average billing amount?

(b) Explain the purpose of the following code options used in the `PROC SGPLOT`:

- `response=Avg_Billing_Amount`
- `categoryorder=respdesc`
- `datalabel`

4. Exploring Admission Types

(a) How does the code ensure that spaces are handled when filtering for 'Urgent' and 'Elective' admissions?

(b) Which procedure is used to check the unique values in the `'Admission Type'` column, and what is the output of this procedure?

5. Comparison: Urgent vs. Elective Admissions

(a) How does the code calculate the average billing amount for different admission types?

(b) What is the purpose of using `strip()` in the following code?

```
if strip('Admission Type'n) in ('Urgent', 'Elective');
```

6. Visualization: Admission Type Comparison

(a) Describe the plot that is used to compare average billing between ‘Urgent‘ and ‘Elective‘ admissions.

(b) Explain the importance of the following options in the bar chart:

- `stat=mean`
- `datalabel`
- `categoryorder=respdesc`

More expansion over question 2:

1. Understanding the Use of Name Literals

(a) Why does the following code use the `'column name'n` syntax for variables like `'Date of Admission'n`?

(b) What would happen if you referenced these variables without the name literal syntax?

```
if 'Date of Admission'n > 0 and 'Discharge Date'n > 0 then
    Length_of_Stay = 'Discharge Date'n - 'Date of Admission'n;
else
    Length_of_Stay = .;
```

2. Calculating Length of Stay

(a) In the code above, how is the **length of stay** calculated for each patient?

(b) Why is the `else` block assigning a missing value (`.`) to `Length_of_Stay`?

(c) What condition must be true for the `Length_of_Stay` to be calculated?

3. Handling Missing Data

(a) What does the expression `if 'Date of Admission'n > 0` ensure in the code?

(b) How does the program handle missing or invalid dates when calculating the length of stay?

4. Filtering the Data for Analysis

The following code filters data to include only `'Urgent'`, `'Elective'`, and `'Emergency'` admissions:

```
data valid_admissions;
    set health_data;
    if strip('Admission Type'n) in ('Urgent', 'Elective', 'Emergency');
run;
```

(a) What is the purpose of the `strip()` function in this code?

(b) Why might it be necessary to use `strip()` on character data when filtering?

5. Calculating the Average Length of Stay

(a) Which **SAS procedure** is used to calculate the **average length of stay** for each admission type?

(b) Write the output dataset created by the following `PROC MEANS` step and describe its structure:

```
proc means data=valid_admissions noprint;
    class 'Admission Type'n;
    var Length_of_Stay;
    output out=avg_length_of_stay mean=Avg_Length_of_Stay;
run;
```

6. Visualizing the Results

(a) What type of plot is used to visualize the average length of stay by admission type?

(b) Explain the purpose of the following options in the `PROC SGPLOT` code:

- `response=Avg_Length_of_Stay`
- `categoryorder=respdesc`
- `datalabel`