Business Intelligence Project

This project is designed to help you to consolidate your learning in the course. You will choose a dataset and explore it to identify and visualize underlying patterns.

Deliverable

Project report as s single file – details later in this document.

Tasks

1. Find a dataset from any source – the dataset should have **at least 500 rows** and at least **two categorical attributes and at least two numerical attributes**. You can search on the web for datasets. You should not use any of the datasets that we have already used in the course materials.
2. Run the dataset by me as soon as possible so that I can see that it is appropriate for the project and provides enough scope for analysis.
3. Identify the categorical and numerical attributes in the dataset (categories and measures).
4. Perform basic analysis on the dataset – like overall summaries, analysis of individual columns, appropriate plots of individual columns depending on the nature of the column, etc. For each analysis, you should show the R code and also include your interpretation of the results. You should include at least five basic analyses and make sure that each one says something significant about the dataset. Just listing the number of rows and columns will not count towards the five, although you can include that in your analysis.
5. Perform more advances analysis – reports and plots that relate multiple attributes. In this section, you will first pose several (at least five) interesting questions that you would like to answer and then provide your answers via reports and charts. You can use the EDA framework that we studied to generate the questions as applicable to your dataset. Generate as many questions as you can and list all questions you generated. Then select those that you think are the interesting ones with reasons for selection. Here are some general themes for generating questions:
    a. two-category counts
    b. distribution of measures with breakup by category
    c. relationships between numeric variables
    d. contribution to measures by individual categories
    e. contribution to measures by two-categories
    f. comparison of variation of measure by category
    g. time series of measures, etc.
6. For each of the above you should show your R code, plots and provide your interpretation. You should make good use of dplyr, pipes and ggplot. Use aesthetics to bring multiple attributes into play and use facets where needed.

Report and data

- Develop the analysis as an **R Notebook** and **generate the pdf** of your report from it.
- Your report should be divided into meaningful sections and your analysis of the results should be written thoughtfully.
- A sloppy report (with spelling and grammatical errors and poor organization, among other things) will cause you to lose points.
- You should not have any output (report or chart) without a subsequent analysis of what it means.
- You should submit only one PDF document with everything included in it.
- Your report should address everything mentioned above. You will lose points if you miss out anything asked for.

Rubric

Category	**Description**	**Points**
Overall structure	The report should have a logical structure with meaningful sections.	10
Basic analysis	Appropriate analysis and interpretation of important attributes	10
Advanced analysis	• Significant list of questions generated and shortlisted. Questions involve multiple attributes where appropriate (10) • Good use of ggplot, dplyr and other packages covered in the course (10) • Meaningful interpretation of reports and charts (10)	30