DATASCI 306, Fall 2025, Final Group Project

Group : mention your group names here

Throughout this course, you've dedicated yourself to refining your analytical abilities using R programming language. These skills are highly coveted in today's job market!

Now, for the semester project, you'll apply your learning to craft a compelling `Data Story` that can enrich your portfolio and impress prospective employers. Collaborating with a team (up to 5 members of your choosing), you'll construct a Data Story using the data provided in the `data` folder. This data is downloaded from: https://data.cdc.gov/browse?sortBy=relevance&pageSize=20&q=Adult+Tobacco+Consumption+In+The+U.S.%2C+2000-Present&page=1

Deliverable

1. Requirement-1 (4 pt)

You should show at least 4 steps you adopt to clean and/or transform the dataset. Some of the steps you might take are; merging all the data into one dataframe, converting datatypes, creating additional columns, cleaning column names etc.

2. Requirement - 2 (20 pt)

You need to plot 10 different diagrams to show correlations, frequencies, and/or relationships between various variables with plots of 5 different types (bar, line, heatmap, facet, etc.). Every plot should have a title and the x/y axis should have legible labels without any label overlaps for full credit. Provide a summary of your interpretations from the plots after each one.

3. Requirement - 3 (5 pt)

By this phase, you have a pretty good understanding of your data. Now, you will apply predictive analytics by building suitable ML models to make some predictions. Ensure you apply the techniques taught in lecture 19/20 to figure out the correct variables to choose, based on p-value, residuals etc..

Build a shiny app that allows the user to input values into the interface and then make predictions using your model. You are required to build the shiny app just for the prediction section of your project. Although we will accept if you create a shiny app for your entire project (not required)

4. Requirement - 4 (1 pt)

You should have a conclusion, highlighting the main insights you were able to derive from your analysis.

This is an open ended project where every team will come up with unique insights. We would like to see what each team comes up with.

Submission

- You will upload the zip file containing your final.Rmd file, final.pdf file of the EDA and app.R file for your shiny app, as a deliverable to Gradescope

- You will present your findings by creating a video of a maximum of 15 minutes duration, explaining the code and the workings of your project; all team members should explain their part in the project to receive credit. You will share the video URL on Canvas for us to watch.

It is not necessary to prepare slides (if you do it doesn't hurt) for the presentation. You may speak by showing the diagrams and/or code from your laptop. Every team member should explain their part in the project along with the insights they derived by explaining the charts and summaries for full credit to each member. An easy way to get this accomplished is to open a Zoom meeting and everyone take turns in explaining while you record the meeting. Add the URL of this recording to Canvas.

Your project will be evaluated for its meaningful/insightful EDA and predictions.

Hints

Some answers you may try to find in this dataset could be:

- Are smoking disparities related to income getting worse or better over time?
- Which racial or ethnic group faces the greatest inequality in smoking rates compared to the majority population?
- Do people with frequent mental distress have a much higher smoking disparity than people with disabilities?
- Which 5 focus groups have the highest average disparity over the entire period?

These questions could just get you started but as an analyst you should hone the skills of asking good questions and this project will get you that practice.

Machine Learning Model Development:

You could build a regression model to predict the DisparityValue. Features may include, Year, DisparityCategory, specific FocusGroup etc. Extract and plot the feature importance to select the features.

Summary

Your summary might answer questions like;

- What were the most significant trends and predictors of smoking disparity?
- Which populations should be prioritized for smoking cessation programs? etc.
  You may also include ideas for future analysis and any limitations you came across with the current dataset