Mar 24, 2025 8:00 PM

Prep for Homework 6

Attendees Caroline Causey Sophia Scherer Shreya Holikatti Alex Creech
Attachments

Agenda

Topic	Time	File	Team member
Review Homework 6 Instructions	10 min	File	Shreya Holik…
Discuss individual data columns	10 min	File	ALL
Next Steps + Conclusion	3 min	File	Caroline Cau…

Summary or key decisions

- Split based on ecomp explanatory variables
  - Option based compensation (most impact?) - Sophia Scherer
  - Stock based compensation - Shreya Holikatti
  - Base salary (least impact?) - Alex Creech
  - Total compensation - Caroline Causey
  - Look at each to the total income tax / total revenue
    - Effective tax rate, deferred tax percentage
    - Keep blue for controls, company statistics for exploration

Action items

- [ ] EDA
- [ ] Make a drive folder for Part 2 with all of our colab notebooks

Details

Raw notes

- Ideas for how to split
    - By independent explanatory variables
        - This personally makes more sense
        - Compensation variables 1,2,3,4
    - By dependent variables
- We don't need to build any models
- We each have a query, basic exploration, plan for analysis and modeling approach

Ideas for later

Instructions

12% of the course grade. Due by 11:59pm on March 31, 2025 (online).

1. Craft questions for explorations: As a team, submit at most 5 page document for a 4 member team and 4 page document for a 3 member team.
    - (4% of grade) Craft and implement 3-4 interesting queries or smaller datasets of the larger dataset that would show different aspects of the data points, visualize them to gain insight.
    - (4%) Perform exploratory data analysis around these queries.
    - (4%) Submit an analysis sketch with a clear indication of goal for the team and sub-goals for each member.
2. GitHub: The same Github link as in Part 1 would be used.
3. Assignment: In at most *1 page per individual (with images)*, each individual in the team will submit a brief on:
    - The querying, exploration and visualization could inform one another.
    - Each team member should explore a different question or aspect of the data.
    - You are not expected to fit a model for exploratory analysis.

- Combine the individual parts with names of each member and individual Colab links in a single document.
- In your writeup, include your preliminary question of interest, the goal (what you want to learn from the data) and explain in 5-7 sentences how the individual questions explored in query and exploration by each member ties together for this goal and what subgoals will be explored by the team.
- Add a description of the approach/method that you would like to try (how you want to learn it) for each subgoal.
- Add details of what the unit of analysis will be for the team as the explore these subgoals.
- Clearly mention the dates and the responsibilities of each member in the analysis sketch for the next few weeks.
- *Data Querying*: Provide the 1 query each and your rationale behind running those query in 3-5 sentences for each. (Paste 2 rows of response from your data with each query), specify how many rows were returned and how much time the querying took as reported by Python Kernel. Complex queries will take longer to execute.
- *Exploratory data analysis and visualization*: Explain the steps with code and comments for exploratory data analysis and visualizing your data. Each page should have at-least 2 visualizations and write 3-5 sentences of how you found this exercise interesting in knowing more about your dataset.
- *Analysis Sketch:* In at most one page (as the last page of the report), the team should write a brief description of the analysis they plan to do.

4. One of the common misconceptions about analytics on large quantities of data is that to analyze such data, it is enough to visualize the data (or some summaries of the data) or to summarize the data, and then to draw conclusions based on the appearance of the visuals or on the summaries.
5. Please make sure that your analysis work in the course project is not along the lines of "analyze the data by discussing its visualizations or summarizations, or by discussing the results of the statistical processing of the data." Use the visualizations and querying to inform your analysis sketch.
6. You are not restricted to methods that you have learned in class, and you may have to do some outside reading to figure out some of the details. But provide good rationale for using the exploration steps and visualization.

7. If you are using generative AI, additional submission is required. Refer to the class AI policy for more details.
8. There is also no restriction on the python libraries you can use.
9. As a team, be sure that each team member has a query, exploration and visualization and an responsibility for the implementation in the next coming weeks that they are confident they can implement – this becomes important for securing the team related grades
10. More generally, you can state any of the supervised and unsupervised learning algorithms that fits your goals in your analysis sketch. But be ready to give a rationale of why that was a good choice for the question you were exploring.
11. For examples of advantages and disadvantages of methods discussed in class refer to the image at : https://www.datacamp.com/cheat-sheet/supervised-machine-learning-cheat-sheet