correlation one

MINI PROJECT I:

EXERCISE ON DATA QUESTIONS & BASIC VISUALIZATIONS

EXERCISE INSTRUCTIONS

BACKGROUND

A key component of good analysis is asking the right questions. Good questions spur thinking, frame problems conceptually, and set you on course to discover the answers you need. *What is the best-selling item?* Is a fundamentally different inquiry than, *What is the best-selling item since the pandemic began?*, with large implications for anyone involved.

Once you frame your questions right, visualizations are a crucial way to not only understand your data, but also convey it to other stakeholders. These stakeholders may not have followed your full journey of analysis. So it is critical to describe to them the answers in rich, fitting visuals that they can quickly understand.

ACTIVITY

In this activity, you will use your design skills as well as your data reasoning skills to come up with the best insights and visualization(s) for a given dataset. The purpose of this exercise is to put you in the driver's seat moving to solve a real-life business problem.

Note that when making a data visualization, it is often preferable to make a simple rather than overly crowded or complex visualization; more bells and whistles are not always better. Keep this in mind as you approach this exercise. As you grow your skills throughout this course, you will learn new tools and strategies that will eventually help you create not just visualizations but full-featured dashboards used by top-performing teams of all kinds. We will cover this skillset later, when we learn about Tableau in the course.

DIRECTIONS

In your Project Group, review the **Options** below. Then pick one and its corresponding dataset – whatever you find most interesting. Together you will work to answer all of the questions in your chosen prompt.

Once you have answered the questions, select one or more of the questions to answer in the form of a visual. For example, if the question is *What is the most popular electric car model by month?,* you might want to make bar chart(s) comparing the models by monthly sales. If the question asks about an event over time, perhaps you should use a line graph or other visual describing durations. It is up to your group.

DELIVERABLE

You should create one or two deliverables for this exercise. The first is the main assignment, the second is optional for advanced teams:

1. A written report answering each of the questions and briefly explaining your process to get there
2. [Optional for advanced teams] A visual or series of visuals to answer one of more the questions

Bonus: pose a question or your own and answer it about the dataset

You will submit your deliverables on the course training website, as two different files:

1. The Written Report will be submitted at "Written Answers to Questions (Mini Project I)". This can be a .docx, .doc, or .pdf file*
2. [Optional for advanced teams] The Excel with accompanying visualizations will be submitted at "Excel – Analysis and Visuals (Mini Project I)" This can be a .xlsx file*

*If you encounter issues with either submission, your team should contact your TAs on Slack for a walkthrough.

NOTE

- Here are some example Exercises: Example 1 & Example 2
- There are **hints** for each question at the bottom of the final page. Consult them if you feel stuck
- This exercise will be graded on **effort**, not correctness or completion, so you have the freedom to experiment and try new things without the fear of getting the questions wrong
- Even if you aren't able to answer every single question, demonstrating an honest effort to answer one or more of the questions will give you full credit for the exercise

EXERCISE OPTIONS

OPTION 1: SUPERSTORE PRODUCTS AND SALES

EXCEL FILE LINK

Business problem: You are a data analyst working for a big box retailer. You're given a dataset of orders from the past few years, and your job is to answer questions and, if you can, create a set of visualizations that can help the execs get a clear picture of the company's sales performance. Some questions that were brought up in the last earnings call include:

1. What are our most popular products?
2. What are the most popular product categories?
3. Which customer segments drive the most sales?
4. In which categories do we generate the most sales?
5. Which regions of the country drive more sales in certain categories?
6. How have our products performed over time?
7. Bonus: pose a question of your own and answer it

Take these questions into consideration as you decide which visualizations the company needs, or create your own to show insights not previously thought of.

OPTION 2: NYC TAXI RIDES

EXCEL FILE LINK DATA SOURCE LINK

Business problem: You are a data analyst for a NYC taxi company, and your job is to analyze a dataset containing records of recent trips and provide the company with insights about its riders. Your questions include:

1. How has the average trip duration changed over time?
2. What are the most popular ridership times during the day?
3. How do distance and time of day affect fare prices?
4. Bonus: pose a question of your own and answer it.

OPTION 3: STREAMING WARS

EXCEL FILE LINK

Business Problem: You are a data analyst for an ad agency that specializes in targeted TV ads for small businesses. As more consumers continue to "cut the cord" and move to online streaming services like Netflix, Hulu, Disney+ and Prime Video for their entertainment content, your company wants you to help them identify which services they should purchase ad space from. Some key questions the executives have asked you include:

1. How many TV shows and movies does each streaming platform offer?
2. Are there any titles (shows or movies) that appear on multiple platforms?
3. Who are the most popular directors of titles across all of these platforms?
4. Which platform has added the most titles in the past 1/5/10 years?
5. What is the average movie length across these platforms? What is the average movie length for each platform?
6. What is the most popular genre/category on each platform?
7. Bonus: pose a question of your own and answer it.

OPTION HINTS

EXERCISE 1

1. Hint: Find the count of all of the unique product IDs
2. Hint: Group the products by categories, then perform a calculation of all sales under those categories
3. Hint: Look at the customer table to find which meaningful categories you can group the sales by
4. Hint: Take the sum of all of the sales grouped by product categories
5. Hint: Look at the Customer table to find which regions the sales are coming from
6. Hint: Follow a subset of products and their respective sales over a given period of months and years.

EXERCISE 2

1. Hint: Subtract pickup time from drop-off time to get the total trip duration
2. Hint: Group the ride times into hours in order to get a more meaningful segment of rides
3. Hint: make a scatterplot of the distance and fare of each trip to see if you can spot any trends

EXERCISE 3

1. Hint: Look at the "type: column in each sheet to calculate the frequencies for each
2. Hint: The "title" column in each dataset will be the best data point to answer this question
3. Hint: The "director" column contains the directors of all titles in the dataset.
4. Hint: The "date_added" column contains the date that the title was added to the platform
5. Hint: The "duration" column contains the runtime of all of the titles in the dataset. Be sure to filter out the TV shows and just calculate the runtime for the movies!
6. Hint: Use the "listed_in" column to find the associated categories of each title