CAREERFOUNDRY

Data Immersion Achievement 6 Project Brief: Advanced Analytics & Dashboard Design

Objective

To build an interactive dashboard that will visually showcase well-curated results of an advanced exploratory analysis conducted in Python.

Context

For this Achievement, you'll select your own data to analyze, with the goal of conducting an exploratory visual analysis in Python and finding connections between variables that seem worth exploring. After developing hypotheses, you'll use various advanced analytical approaches to help you test your hypotheses.

The results of your analyses will be presented in a Tableau dashboard/storyboard. Data dashboards are an effective tool for, among other purposes, presenting data in an accessible and tangible way. Your dashboard will tell the story of your analytical journey and, as such, needs to contain a curation of the key results you discovered throughout this Achievement.

Note that not all of your results will fit into the dashboard. Any additional analyses you conducted as part of the project will need to be included in a GitHub repository.

Data Requirements

As explained above, you'll be **sourcing your own data** for this project. However, in order to conduct the procedures explored in this Achievement (and develop the necessary skills for a junior analyst), the data you choose will need to meet some specific criteria. You'll likely need to source more than one data set while working through the Achievement in order to meet the different criteria. It is, however, sufficient to start out with one that will act as your main data set. Keep in mind that data sourcing could be very time consuming!

The data set(s) you choose must:

- Be open-source.
- Come from an authentic/authoritative source.
- Include non-anonymized column names.

- Be recent (ideally, no more than 3 years old. However, this factor is not essential - if you've found a perfect data set for your purposes, it could be older too, but not more than 10 years old).
- Contain at least 2-3 continuous variables (apart from index variables, ID variables, dates, years, etc).
- Contain at least 2-3 categorical variables (apart from index variables, ID variables, dates, years etc).
- Contain at least 1,500 rows.
- Include a geographical object of some kind: for instance, a column relating to a country, continent, or something similar. If the information from the data set refers to the US, for example, there should be a column containing the names/abbreviations of the states. Note: there should be at least a couple of different values in this column. This is important for the geospatial analysis you'll be conducting. Of course, you can also use data sets with latitude and longitude.
- In the course of the Achievement you will source a time series data set too, but this procedure will be explained explicitly in the corresponding Exercise, so you don't need to worry about it now.

In summary, the **__most important feature__** your data set(s) need to contain is a **__geographic feature__**. This means the data you collect should refer to countries, states, towns and so on - generally, anything you can visualize on a map when conducting geospatial analysis. In Exercise 6.1, you'll find some suggestions for the type of data to use for this project based on your professional experience and goals.

If you need data sourcing is taking too much time or you are unable to find a viable data set, you can choose one of the following data sets for this Achievement:

- Boat Sales
- New York Citi Bikes
- World Happiness Report 2015-2019
- Airbnb Amsterdam
- Brazilian E-Commerce
- House Sales in King County, USA
- Zomato Bangalore Restaurants to be used with this JSON
- Medical Cost Personal Datasets to be used with this JSON
- World University Rankings
- Chocolate Bar Ratings
- Gun Violence Data
- UFC-Fight historical data from 1993 to 2021

Analysis Criteria

You'll be conducting the following analyses in this Achievement (note that not all results need to be included in the final dashboard):

- Exploratory analysis through visualizations (scatterplots, correlation heatmaps, pair plots and categorical plots)
- Geospatial analysis using a shapefile
- Regression analysis
- Cluster analysis
- Time-series analysis
- Analysis narrative and final results (presented in a dashboard)

Dashboard Requirements

As explained above, your dashboard needs to tell the story of your analytical journey. As such, you'll want it to contain a curation of the key results you discovered as you worked through the Achievement. Based on your findings, you can decide which visuals and procedures are useful to include. (Note: Anything you don't include in your dashboard can still be included in your GitHub repository.)

Your final dashboard must:

- Be designed with a use-case in mind (answering key guidance questions).
- Be created in Tableau Public.
- Be interactive.
- Adhere to visual design best practices.
- Include an introduction page that describes the project (data and purpose).
- Include relevant result(s) of initial visual exploratory analysis.
- Include an explanation for how the results of the exploratory analysis resulted in defining research questions and/or hypotheses.
- Contain a geospatial component.
- Address the defined questions/hypotheses using advanced analytical techniques. For example:
    - Regression analysis
    - Cluster analysis
    - Time-series analysis
- Include a results summary page explaining how the results do or don't address the initial research questions/hypotheses.
- Include details on the limitations of the project.
- Include a proposal of next steps for further analysis.

GitHub Repository Requirements

Your dashboard will be created in Tableau, but you'll also need to ensure the analysis you conducted in Python is available for viewers, as well. This will give you a place to include any

steps of your analysis that don't fit into the narrative of your dashboard. Your GitHub repository must include:

- Your Python code.
- A logical folder structure.
- Folders and files that have been named following industry-standard conventions.
- Portfolio-ready Jupyter scripts for every task in the Achievement (complete with code comments, organized structure, and clean, functioning code).
- A README file that contains a description of the project, details of the data source(s), research questions, and cleaning procedures (from Exercise 6.1 task), as well as a link to the Tableau dashboard containing the analysis results and a short description of its contents.

Your Project Deliverables

Throughout this Achievement, you'll be working from Exercise to Exercise to complete your project. For each Exercise task, you'll submit a deliverable that directly contributes to the final product—in this case, a data dashboard. Since you can select your own data for this project, you'll begin by conducting an exploration of the data before defining the scope of your project, moving on to your analysis, then finally building your dashboard.

Below is a breakdown of your course project deliverables by Exercise.

Exercise 6.1: Sourcing Open Data

- Source the data you've chosen for your project by adhering to the requirements stated in the project brief.
- Complete preparatory steps before moving on to analysis (e.g., cleaning).
- Define questions to explore based on your understanding of what the data contains.
- Create a document containing details of initial preparatory steps conducted.

Exercise 6.2: Exploring Relationships

- Conduct exploratory visual analysis using relevant Python libraries.
- Use the questions you defined in the previous task to guide your exploration.
- If possible, define hypotheses to test.

Exercise 6.3: Geographical Visualizations with Python

- Source a shapefile containing location data that corresponds to the location data in your main project data set.
- Wrangle, clean, and merge data files in preparation for analysis.
- Conduct a geospatial analysis by creating a choropleth map using relevant Python libraries.

Exercise 6.4: Supervised Machine Learning: Regression

- State your hypothesis.
- Select the relevant variables.
- Prepare your data for a regression analysis.
- Split the data into two sets: a training set and a test set.
- Run a linear regression on the data and analyze the model performance statistics.

Exercise 6.5: Unsupervised Machine Learning: Clustering

- Prepare your data for a cluster analysis.
- Use the elbow technique to determine the optimal number of clusters.
- Run the k-means algorithm.
- Attach a new column to your dataframe with the resulting clusters.
- Create a variety of different visualizations using your clustered data.
- Calculate the descriptive statistics for your clusters using the `groupby()` function and discuss your findings and any proposed next steps.

Exercise 6.6: Sourcing & Analyzing Time Series Data

- Source time-series data relevant to your project data via an API.
- Subset your data if necessary so that it contains only relevant historical data.
- Visualize the data in a line plot and decompose its structure.
- Conduct a Dickey-Fuller test and plot autocorrelations to test for stationarity.
- Perform differencing to stationarize non-stationary data.

Exercise 6.7: Creating Data Dashboards

- Define the use-case for your dashboard.
- Outline dashboard contents based on curated results of analysis.
- Create dashboard/storyboard in Tableau per the requirements in the project brief.
- Publish your storyboard to Tableau Public.
- Create a portfolio-ready GitHub repository for your project per the requirements above in the project brief.

CAREERFOUNDRY