Data Analytics Final: **Data-Driven Decision Making**

For your final, you will utilize data analytics to address a question or assist in making an informed decision. Questions you can answer can range from, “are there any anomalies in this cost data?” or “I was dealt a 20 in blackjack and the dealer shows a 9, can I predict the odds of winning this hand?” or “My boss asked me to identify the months a person may be more predisposed to buying x,y,z”.

The project involves several steps, but starts by asking a question that can be answered by data. Once you know what you are asking/trying to determine you will need to acquire the data. Data acquisition will aid in answering the question you pose. You will then clean the data if applicable. Once the data is ready, do some descriptive and predictive analysis, then visualize the data. Finally, you will creare a concise summary of the findings and their implications, then present to the class (5 min).

Introduction (10%)

- Define a clear question or decision-making scenario that can be addressed through data analytics.

  Using machine learning anomaly detection techniques, can we identify and study extreme minimum and maximum temperatures in London from 2000 to 2020 in comparison to the baselines analyzed from 1980 to 2000 with respect to month.

- Justify the significance of the chosen question or decision-making scenario in a real-world context.

  This is interesting because it is consensus in the scientific community that temperatures globally are rising. So it is relevant to study at a localized level.

Data Acquisition (15%)

- Identify and acquire a relevant dataset from an online source.

  https://www.kaggle.com/datasets/emmanuelfwerr/london-weather-data

- Import the dataset into a Snowflake database, detailing the steps taken.

  Downloaded the .csv file from Kaggle and then imported that .csv file into snowflake using its build in function.

Data Cleaning and Preparation (10%)

- Assess the dataset for quality and cleanliness, does any cleaning need to be done? Any nulls? Outliers?

  There was some cleaning that needed to be done. There were no nulls or outliers that needed to be considered.

- Perform necessary data cleaning and preprocessing steps, documenting the rationale and methods used.

  The transformation that were done is the date column was cleaned and formatted as a time stamp to allow for later date manipulation. The dates were filtered to be within the start of 1980 to the end of 2020. The data was then split between training and test data based on whether it was between the years 1980 to 2000 (train) and the years 2000 to 2020 (test).

Descriptive Analysis (15%)

- Conduct a descriptive analysis to summarize the main characteristics of the data

  Min Temp: -6.2
  Max Temp: 37.9
  Mode: 11.1
  AVG: 15.388
  Median: 15
  Standard Deviation: 6.5547

- Identify patterns, anomalies, or insights that could inform the subsequent predictive analysis.

  A general pattern in the data is it seems temperature is rising throughout the years.

Predictive Analysis using Snowflake Cortex ML Functions (20%)

- Select and justify a machine learning approach (Anomaly Detection, Prediction, or Classification) relevant to the project objective.
  I choose to do an Anomaly Detection approach. In summary it is reasonable to assume that studying a 20 year period of historical weather data, should allow me to train a model to observe if maximum and minimum daily temperatures are anomalies in comparison to thresholds the models learns in the training data. We can apply this model to the test data to see if current weather trends are full of what would be defined

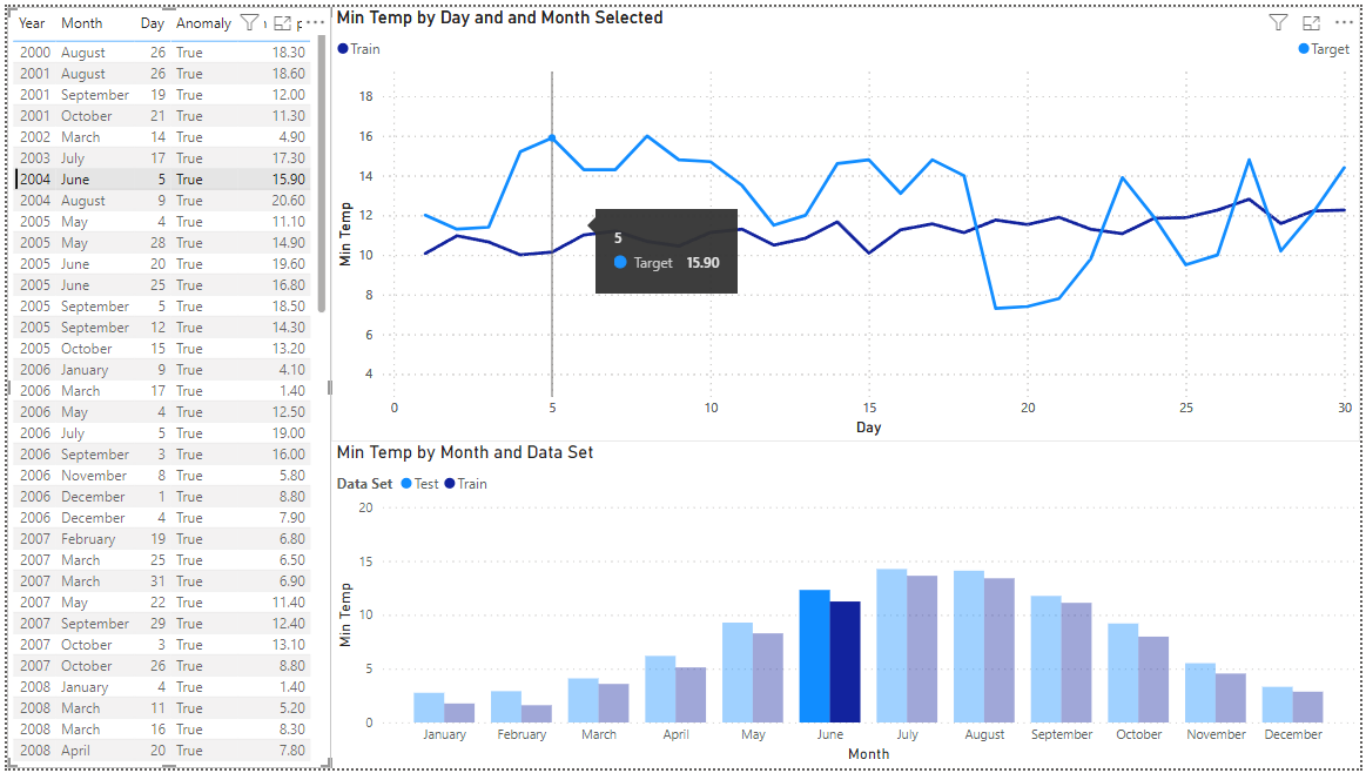
historically as anomalies.

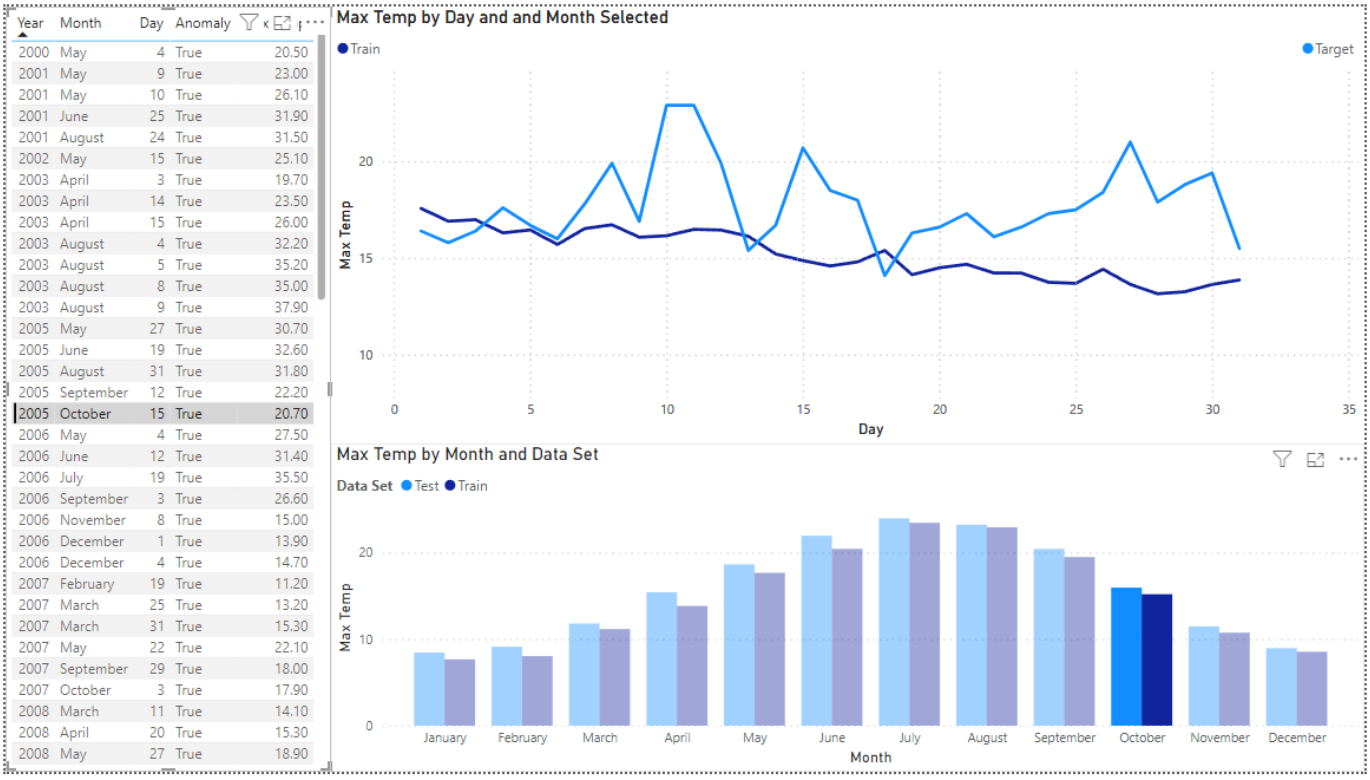
- Train a model using Snowflake Cortex ML Functions, explaining the process and choice of parameters.
  I used the anomaly detection function I split the data to 1980 to 2000 to act as my training data to which my model learned the timed series anomaly detection system. The parameters used to identify if the min temperature of the day and the max temperature of the day were anomalies were the date column and the month column. The test data was the partition of data between 2000 and 2020 where in when I applied the anomaly detection model to identify its stream minimum and maximum daily temperatures.

See https://docs.snowflake.com/en/guides-overview-ml-functions

Data Visualization with Power BI (10%)

- Connect your dataset(s) to Power BI to and visualize your findings.
- Ensure that the visualizations effectively communicate the insights and support the data-driven decision-making process. Below, paste a screenshot of yourvisualization and the question it answers/decision it drives.





- Write a 250-word overview summarizing the key findings, their implications for the question or decision at hand, and potential future work. Include the conclusion below and in your powerbi report.

  To summarize the key findings, I set out to explore information about London weather using machine learning anomaly detection techniques, can I identify and study extreme minimum and maximum temperatures in London from 2000 to 2020 in comparison to the baselines analyzed from 1980 to 2000 with respect to month. Using this anomaly detection system I was able to find that there were a large amount of outliers when comparing the train data which was 1980 to 2000 to the test data which was 2000 2020. So many outliers to assume the fact that are weather is changing in the world. There were multiple occurrences of months in the train data avg temperatures being higher than what the test data months had which further supports the statement that our weather has changed from how it has been in the past. This was not the initial question I was trying to answer, but after looking at the results I found it very interesting to explore more because these results also support a lot of the scientific research that is done today on global warming and how that is effecting our weather. So in conclusion, I found that the data I analyzed supports the theory that the weather that was present in the years 1980 to 2000 is significantly lower in both minimum temperature and maximum temperature than the years 2000 to 2020. I included data visualizations in my power-bi report to help better understand this data and how it looks from a visual perspective.

- You will present your PowerBI report and findings to the class. Ensure you can navigate the report and explain the why and how behind it.

Deliverables

- A completed final report
    - Submit to Canvas
- A link to your repo in GitHub. Use it to store your code, powerbi report, and findings
    - Submit to Canvas
- A Power BI dashboard showcasing the data visualization(s) and findings/conclusions. Upload the file to your github repo
    - Submit to Canvas
- A brief in-class presentation summarizing the project findings and implications
    - In class 4/25

Assessment Criteria

- Clarity and relevance of the question or decision-making scenario.
- Technical proficiency in data handling, analysis, and visualization.
- Logical flow and coherence of the data analysis process.
- Depth of insight and analysis in the predictive modeling.
- Quality and effectiveness of the data visualizations.
- Conciseness and clarity of the final overview.

Deliverables