
AIRBNB-NEW YORK
"RARE FIND"

TEAM : APEX

SINDHURA .ALLA

Sindhura worked on the data preprocessing phase, which involved data cleaning, transformation, and feature selection. She handled missing data, identified and removed outliers, and performed feature engineering to create new variables. She also conducted exploratory data analysis to gain insights into the data and select the most relevant features for the models. Sindhura's contributions were crucial in ensuring that the data used for modeling was clean, accurate, and relevant.

MURALIDHAR REDDY .REDDEM

Murali worked on building the MLP and Decision tree models using Pyspark. He implemented and trained the models, selecting appropriate variables to optimize their performance. He helped to create a dataflow and identify metrics for evaluating the model and selecting the models to achieve a good predictive model for New York Rare Find listing. Murali's contributions were essential in building robust and accurate models that could predict Airbnb rental prices with decent Accuracy.

ASWITH REDDY .KOVVURI

Aswith worked on building the linear regression model using Py-spark. He performed the necessary data transformations, splitting the data into training and test sets and tuning the hyperparameters to optimize the model's performance. He also evaluated the model's performance using metrics such as root mean squared error and mean absolute error. Aswith's contributions were instrumental in building a strong linear regression model that could accurately predict Airbnb rental prices.

BUSINESS PROBLEM

- The business problem/challenge is understanding the factors that affect the price of an Airbnb listing in the New York area. This is important for both hosts and guests.
  - To attract guests and make a profit, hosts need to set the right price for their listing, and guests want to find a listing that fits their needs and budget.
  - Predominantly, the cost of an Airbnb posting is impacted by many variables, for example, neighborhood, property type, room type, conveniences, and host qualities. As a result, determining the best price for a listing can be challenging.
- The motivation for this problem is that Airbnb is a popular platform for travelers to find affordable and unique accommodations and for hosts to earn extra income by renting out their properties. However, setting the right price is crucial for both parties. If hosts overprice their listings, they may not attract enough guests, and their listings may remain vacant. On the other hand, if hosts underprice their listings, they may not earn enough profit, and guests may perceive their listings as low-quality. Moreover, guests may have different budgets and preferences, so they need to find a listing that matches their needs.

Why New York

- New York City is a popular tourism, business, and cultural destination. It is one of the world's most vibrant and diverse cities, attracting millions of visitors annually. As a result, the demand for Airbnb listings in New York City is high, and the market is competitive. This makes New York City an ideal location for a data analysis/modeling project on Airbnb listings and their prices.
- Furthermore, New York City has diverse neighborhoods and property types, from luxury apartments in Manhattan to cozy studios in Brooklyn. This diversity provides a rich dataset to analyze and model, allowing us to explore the different factors that affect the price of an Airbnb listing in different contexts.

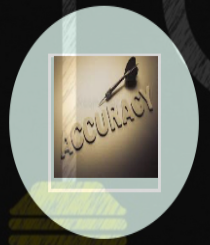
How Data Analysis helps

- Data analysis can address the challenge of understanding the factors that affect the price of an Airbnb listing in the New York area by providing insights into the relationships between different variables and their impact on the price. Here's how data analysis can help:
- *Identifying key factors*: The key factors that affect the price of an Airbnb listing in the New York area, such as neighborhood, property type, room type, amenities, host qualities, and reviews, could be identified.
- *Evaluating relationships*: Relationships between these factors and the price of a listing could be evaluated. For example, it can determine which neighborhoods have higher prices and which property types are more expensive.
- Providing insights: Insights could be provided into the factors that hosts can manipulate to increase the price of their listing and the factors that guests should consider when choosing a listing that fits their needs and budget.
- *Identifying trends*: identifying trends in pricing over time, such as seasonal variations or changes in response to external events.
- *Support decision-making*: Analysis could support decision-making for both hosts and guests by providing actionable insights based on the data. For example, hosts can use the insights to set the right price for their listing, and guests can use the insights to find a listing that fits their needs and budget.

Evaluating Metrics

Mean Squared Error (MSE): MSE measures the average squared difference between the predicted and actual values is used to evaluate the accuracy of predictions.

R-squared (R2) Score: The R2 score measures the proportion of variance in the dependent variable (price in this case) explained by the independent variables. It is used in our regression model to evaluate the goodness of fit.



Accuracy: Accuracy measures the proportion of correct predictions a classification model makes. It is used in our classification models to evaluate the model's overall performance.

Precision and Recall: Precision measures the proportion of true positive predictions among all positive predictions, while recall measures the proportion of true positive predictions among all actual positive cases. These metrics are used in our model to evaluate the model's ability to identify positive cases correctly.

F1-score: F1-score is the harmonic mean of precision and recall. It is used in our model to evaluate the model's overall performance.

DATA SET DETAILS

Listings, including full descriptions, Host details, Availability, Beds, Bathroom count, Property type, review scores, Reviews, unique id for each listing, the price, accommodation, and amenities count.

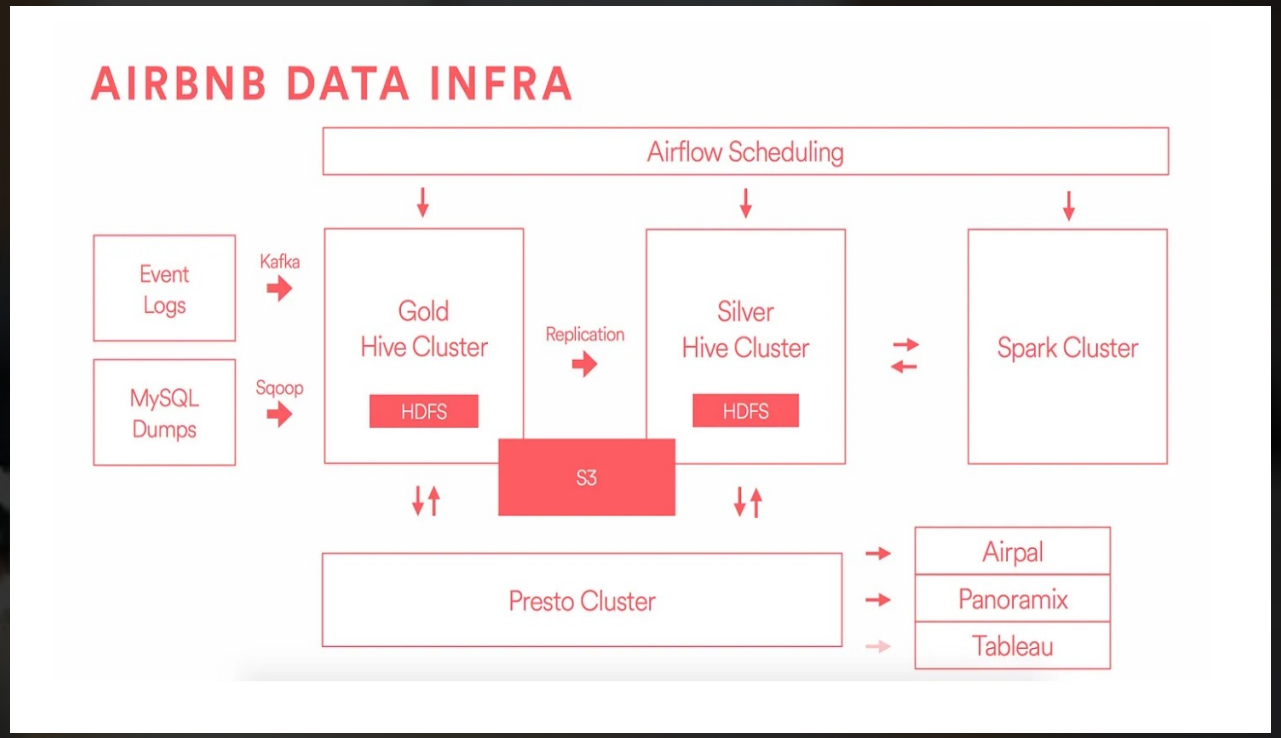
About the Source:

- Airbnb, Inc is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals and tourism activities. Based in San Francisco, California, the platform is accessible via a website and mobile app. Airbnb does not own any listed properties; instead, it profits by receiving a commission from each booking. The company was founded in 2008.
- Airbnb is a shortened version of its original name, Air Bed and Breakfast.com. Inside Airbnb is a mission-driven project that provides data and advocacy about Airbnb's impact on residential communities, where data and information empower communities to understand, decide and control the role of renting residential homes to tourists.

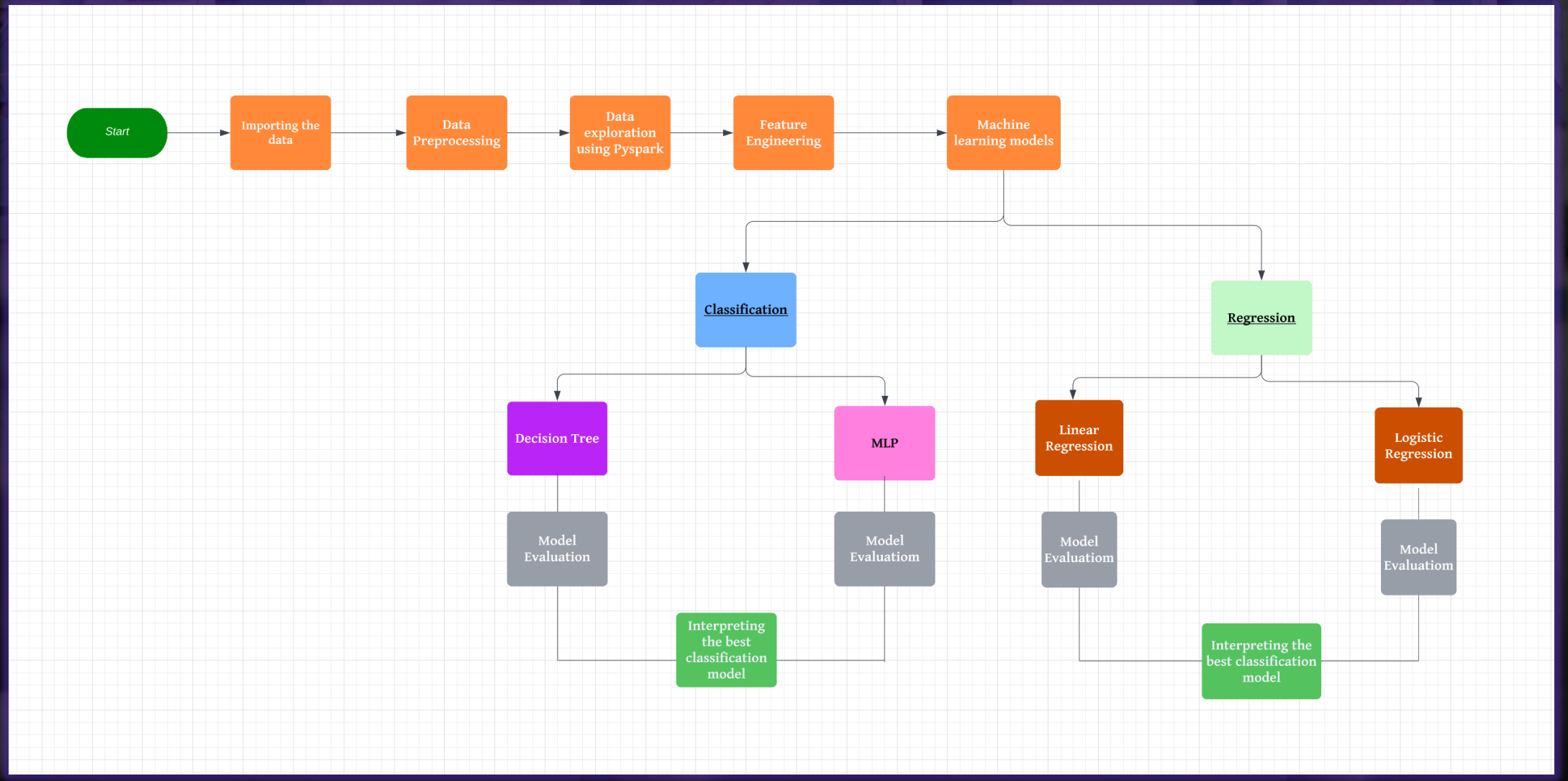
Data source link: http://insideairbnb.com/new-york-city

DATA PROVENANCE

- This flow describes the infrastructure stack used by a company to manage and analyze user activity data.
- The data comes from two channels: instrumentation in source code and production database dumps.
- The data is stored in a "Gold" cluster where extraction, transformation, and load jobs are run. A "Silver" cluster also exists where less critical jobs can run.
- The company uses HDFS and Hive-managed tables as a central source and sink for data, discouraging the use of separate infrastructure.
- Ad hoc queries are run using Presto, and scheduling through Airflow.
- The Spark cluster is used for machine learning, and data can be retired to S3 for long-term storage.

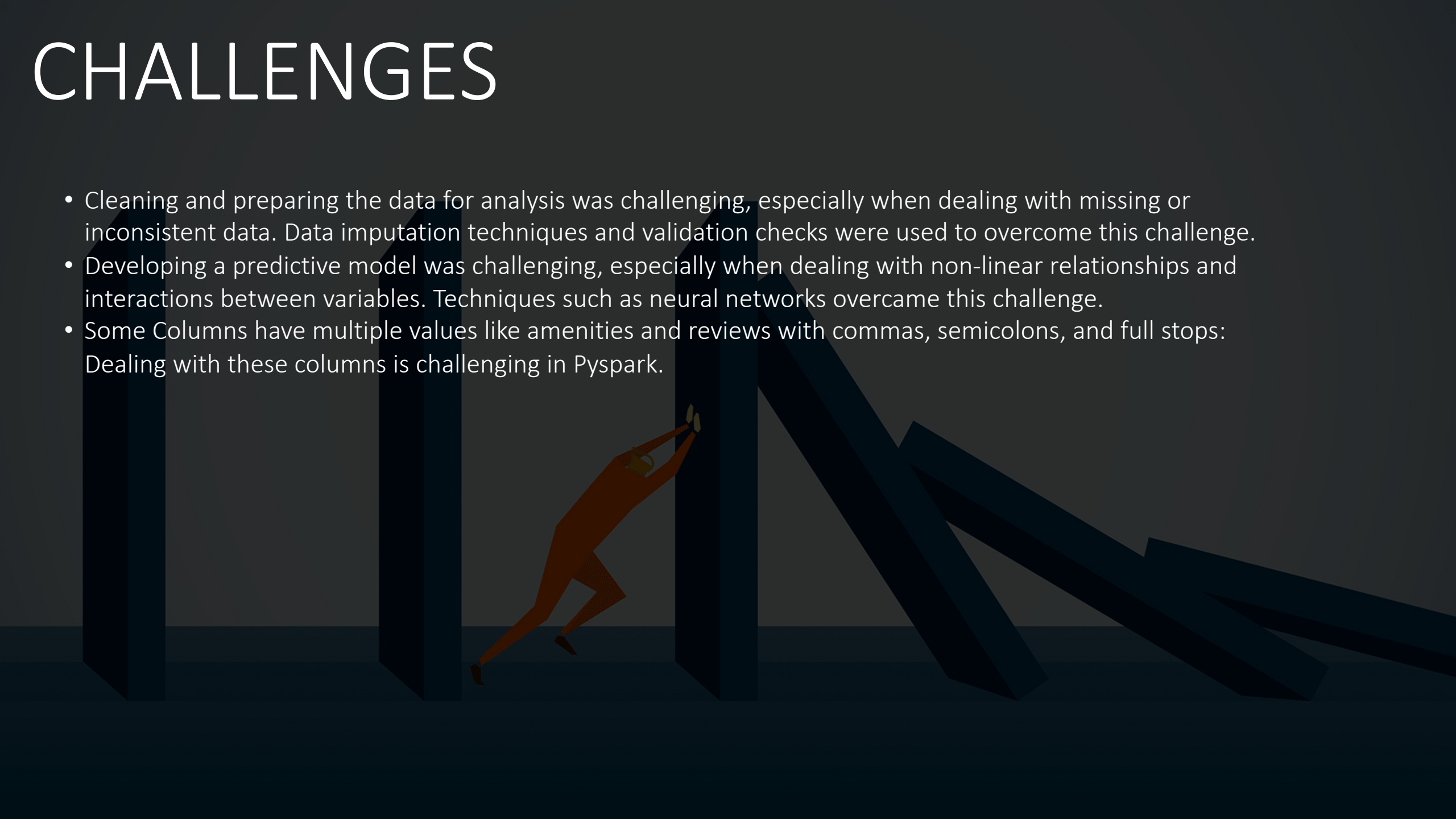


DATA FLOW MODEL



CHALLENGES

- Cleaning and preparing the data for analysis was challenging, especially when dealing with missing or inconsistent data. Data imputation techniques and validation checks were used to overcome this challenge.
- Developing a predictive model was challenging, especially when dealing with non-linear relationships and interactions between variables. Techniques such as neural networks overcame this challenge.
- Some Columns have multiple values like amenities and reviews with commas, semicolons, and full stops: Dealing with these columns is challenging in Pyspark.



REFERENCES

- https://medium.com/airbnb-engineering/data-infrastructure-at-airbnb-8adfb34f169c

THANK YOU