Capstone Project

Luis Morales

Luis Morales

- Data Engineer, ML and AI enthusiast
- Drummer at @darleofficial
- Dogs person
- Coffee lover
- LinkedIn: luis-morales-ponce/

Important Notes

Identify yourself in Zoom, using your name and last name

Mute your microphone along the course

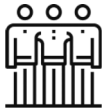
Use the chat for questions during the Q&A sections

Focus your questions on the presented topic

Turn off your camera in case of connection issues

Academy Code of Conduct

Be respectful, there are no bad questions or ideas.



Be welcoming and patient

Be careful in the words that you choose

Session Goal

At the end of this session, you will be able to:

- Identify the main technologies used to reach the data-engineering capstone project.
- Recognize the steps to build all layers.
- Answer some important questions about the data.
- Determine next steps.

Agenda

I. Introduction

Understand the main objective to reach in the capstone project.

II. Infrastructure and Architecture

Identify the principal technologies used.

III. Airflow DAGs

Analyze the DAG implementation and it's logic

IV. Analytics

Answer specific business questions and more...

V. Next Steps

Introduction

Create an **end-to-end** data solution throughout a **Cloud Service** to answer specific questions and make analytics.

This capstone project can be splitted into 3 main parts:

- Build an **Airflow** Cluster in the cloud
- Manage databases though Airflow
- Construct an **ETL** pipeline implementation
- Make analytics

Agenda

I. Introduction

Understand the main objective to reach in the capstone project.

II. Infrastructure and Architecture

Identify the principal technologies used.

III. Airflow DAGs

Analyze the DAG implementation and it's logic.

IV. Analytics

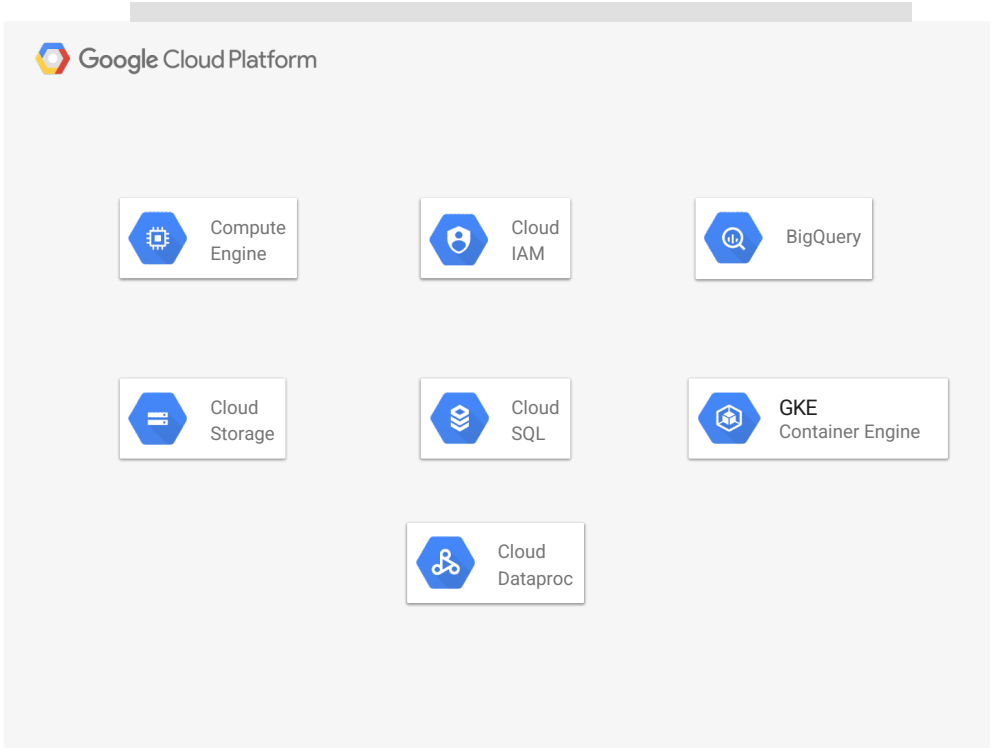
Answer specific business questions and more...

V. Next Steps

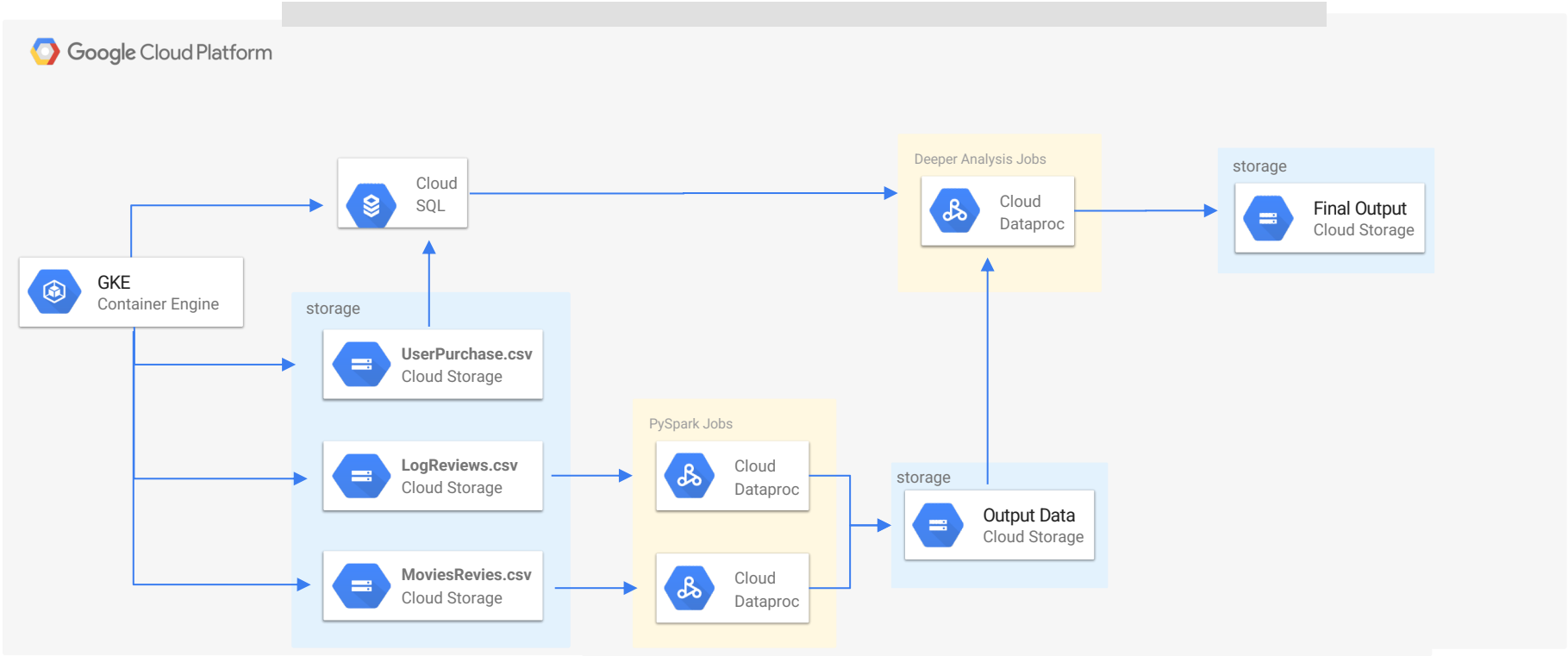
Infrastructure

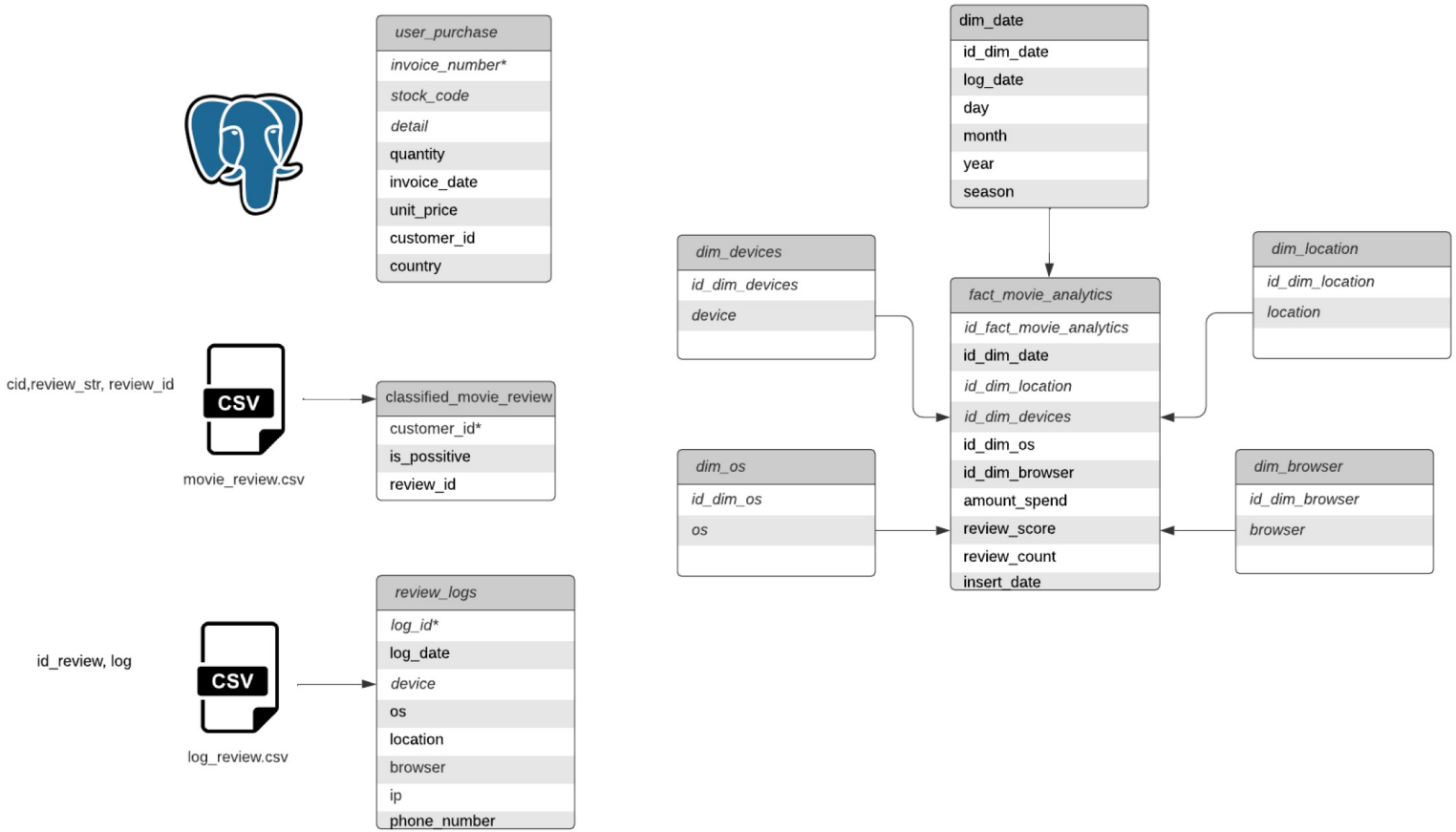
Technologies and Services

- GKE
- Service Accounts
- Cloud SQL
- Cloud Storage
- Airflow
- Dataproc
- Terraform
- Among others



Architecture Overview




user_purchase
invoice_number*
stock_code
detail
quantity
invoice_date
unit_price
customer_id
country
cid,review_str, review_id
CSV
movie_review.csv
classified_movie_review
customer_id*
is_possitive
review_id
id_review, log
CSV
log_review.csv
review_logs
log_id*
log_date
device
os
location
browser
ip
phone_number
dim_date
id_dim_date
log_date
day
month
year
season
dim_devices
id_dim_devices
device
fact_movie_analytics
id_fact_movie_analytics
id_dim_date
id_dim_location
id_dim_devices
id_dim_os
id_dim_browser
amount_spend
review_score
review_count
insert_date
dim_location
id_dim_location
location
dim_os
id_dim_os
os
dim_browser
id_dim_browser
browser

Agenda

I. Introduction

Understand the main objective to reach in the capstone project.

II. Infrastructure and Architecture

Identify the principal technologies used.

III. Airflow DAG

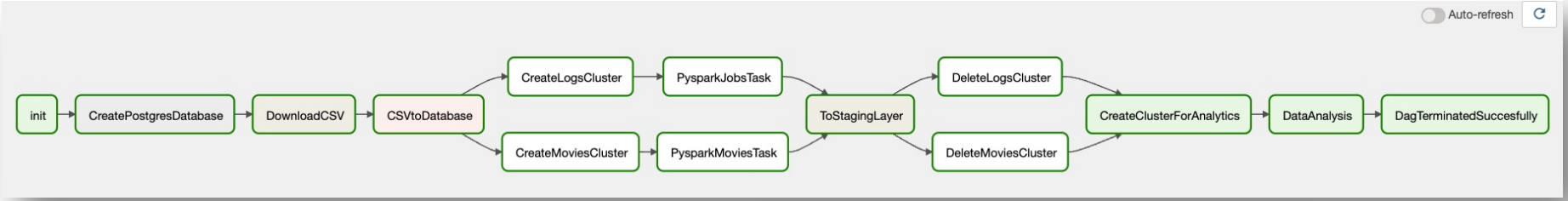
Analyze the DAG implementation and it's logic.

IV. Analytics

Answer specific business questions and more...

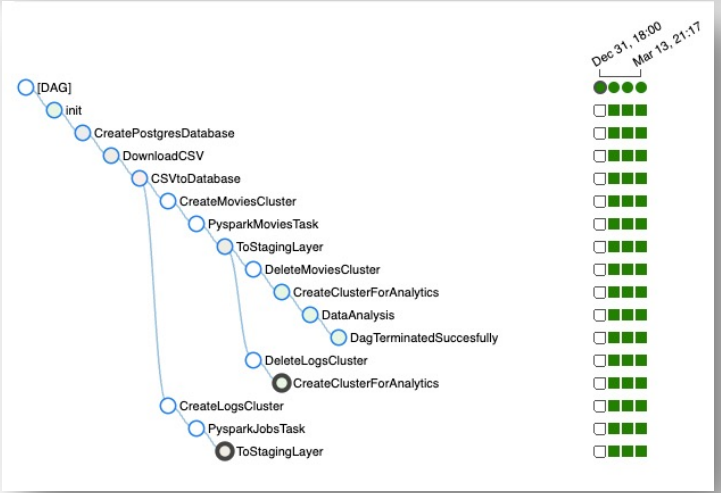
V. Next Steps

Airflow Dag



Usage of:

- Jinja Templates
- GCS operators
- Postgres Operator



Agenda

I. Introduction

Understand the main objective to reach in the capstone project.

II. Infrastructure and Architecture

Identify the principal technologies used.

III. Airflow DAG

Analyze the DAG implementation and it's logic.

IV. Analytics

Answer specific business questions and more...

V. Next Steps

Analytics

- How many reviews were done in California, NY and Texas?

location	count
California	2057
New York	1957
Texas	2015

- How many reviews were done in California, NY, and Texas with an apple device? And how many for each device type?

location	device	Count
California	Computer	698
	Mobile	703
	Tablet	656
New York	Computer	640
	Mobile	661
	Tablet	656
Texas	Computer	638
	Mobile	650
	Tablet	727

- Which location has more reviews from a computer in a Chrome browser?

location	Count - Chrome Browser
Massachussets	159
Montana	156
South Dakota	154
Nevada	151
Washington	148

Analytics

- Which device is the most used to write reviews in the east and which one in the west?

device	Count - West
Mobile	8853
Tablet	8680
Computer	8657

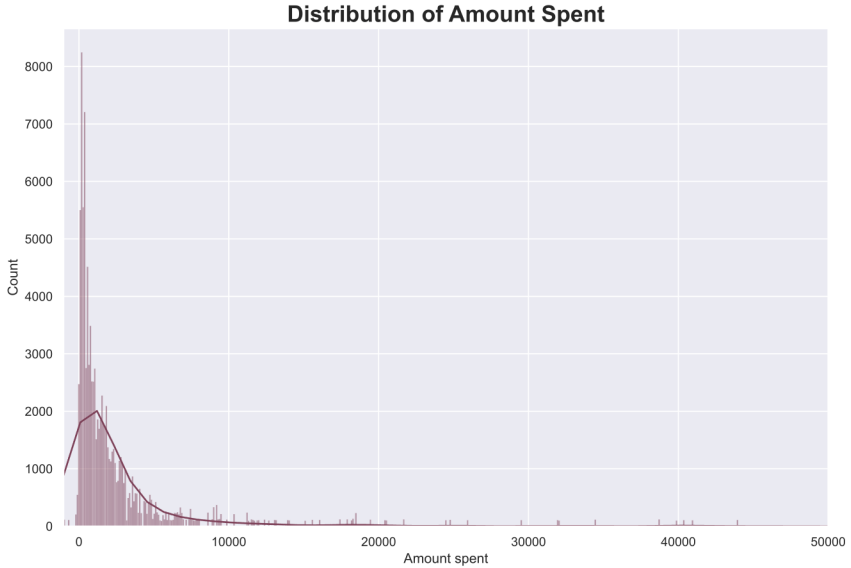
device	Count - East
Mobile	7347
Tablet	7347
Computer	7289

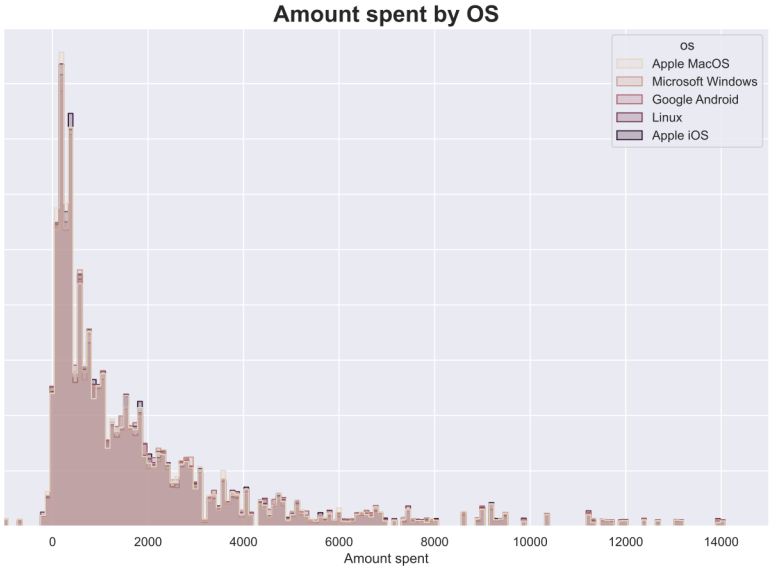
- What are the states with more and fewer reviews in 2021?

location	Count
Georgia	2100
Vermont	1888

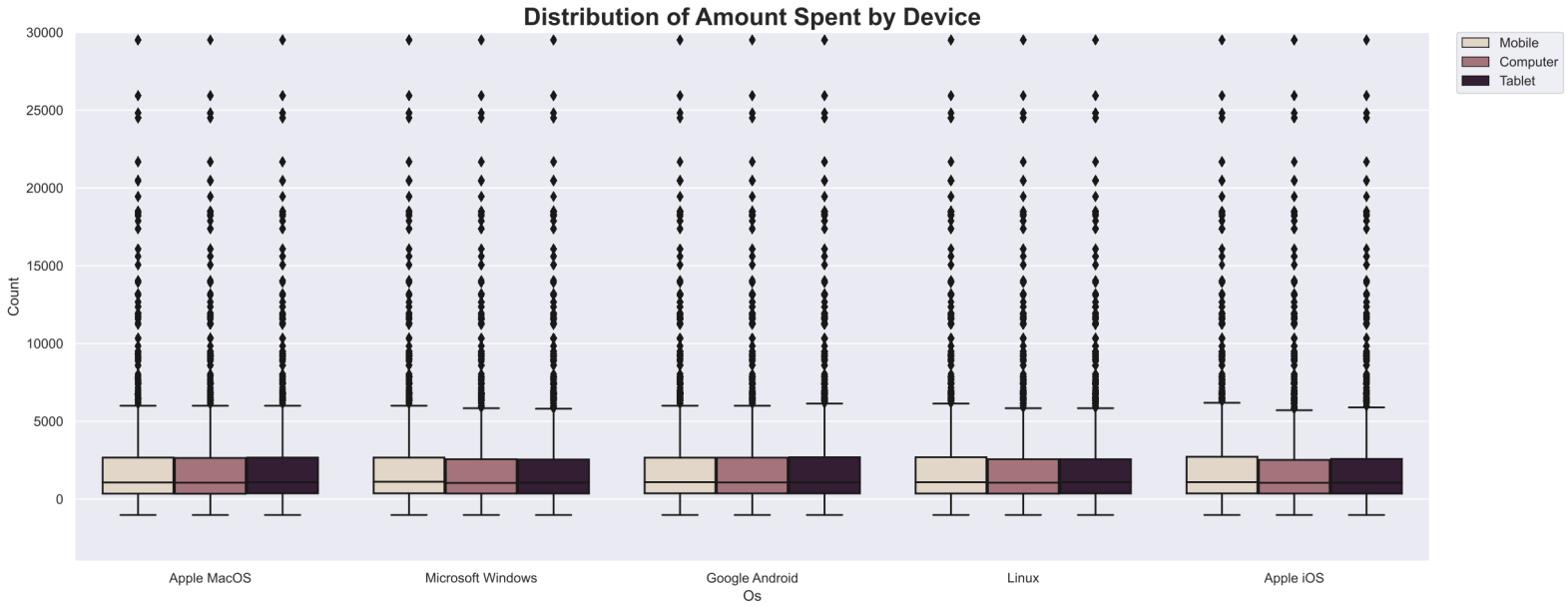
More Analytics

Distribution of variables

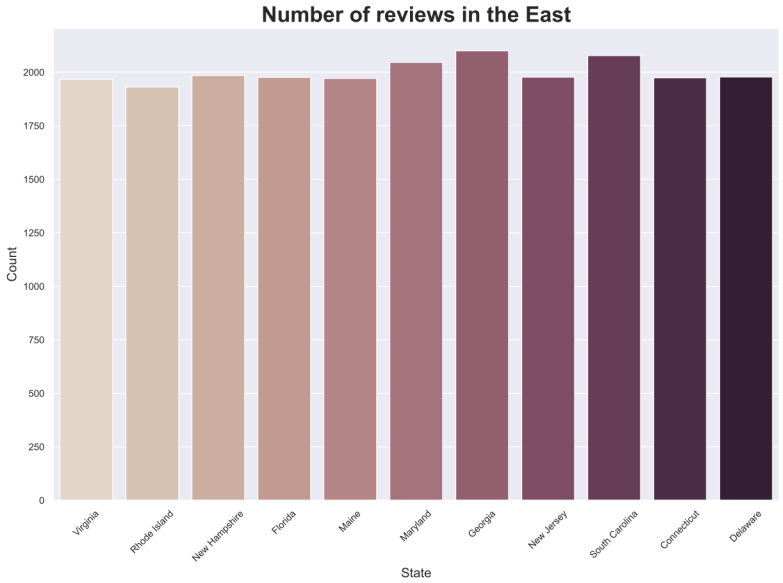


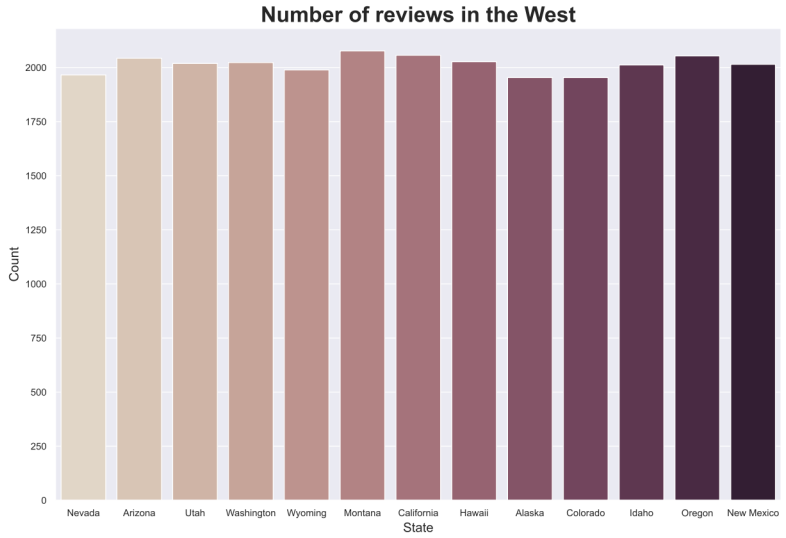


Distribution of Variables

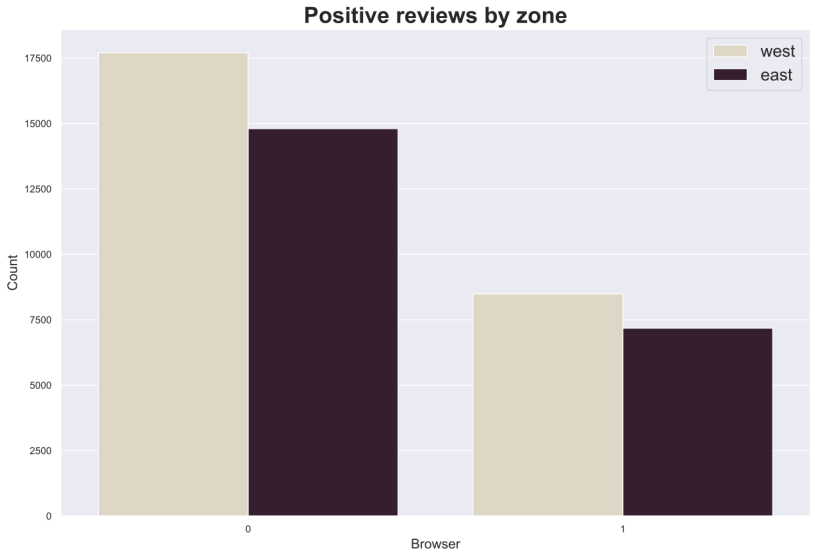


Reviews by State





Positives Reviews

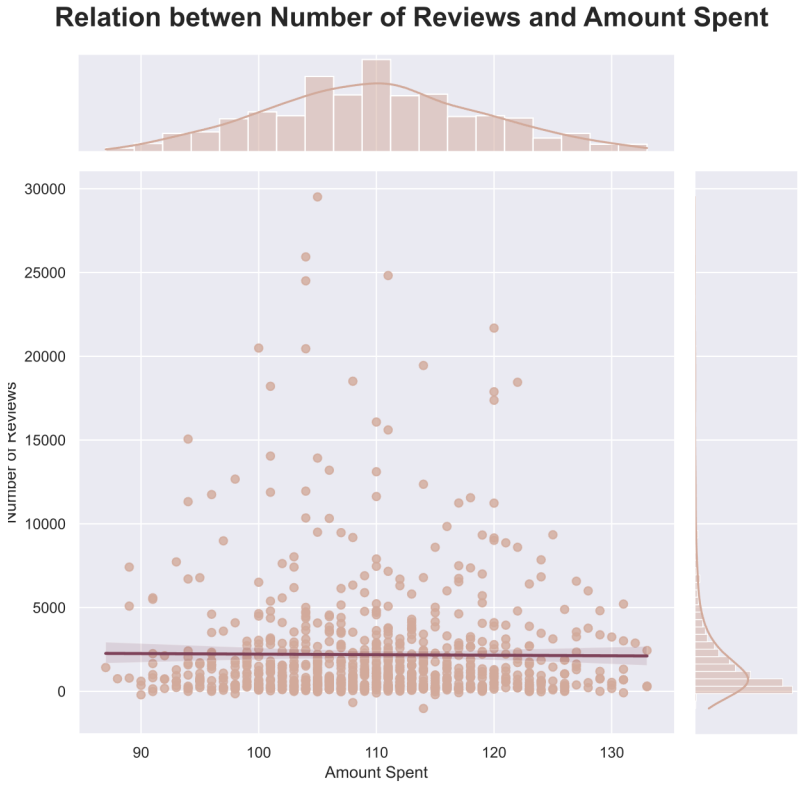


zone	positive_review	count
east	0	14803
	1	7180
west	0	17703
	1	8487

36%

32%

Regression models and outlier removal



Mahalanobis Distance

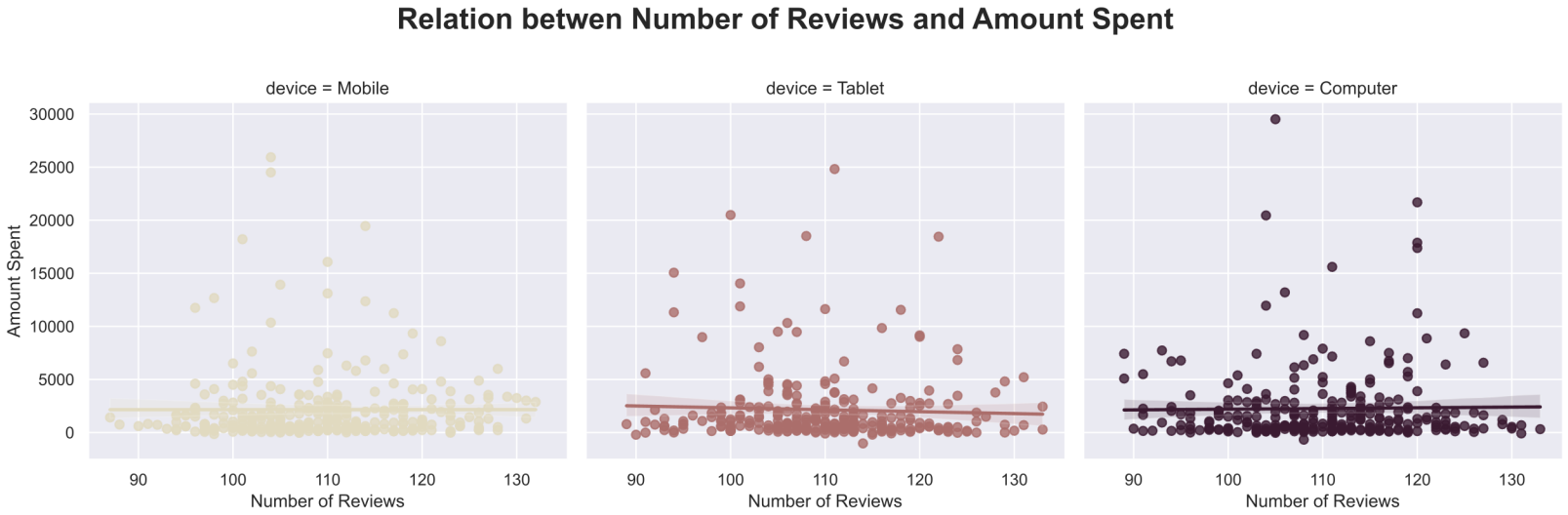
$$d_m(\vec{u},\vec{v}) = \sqrt{(\vec{u} - \vec{v})^T \Sigma^{-1} (\vec{u} - \vec{v})}$$

$\vec{u}, \vec{v}$ *vectors*

Σ *the covariance matrix*

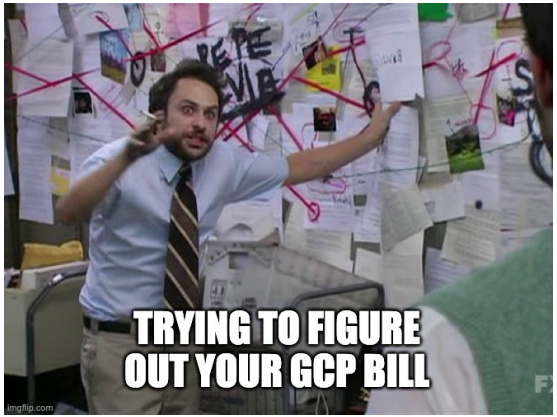
χ^2 *test to determine statistical significance*

Regression models



Lessons learned

- GCP Serverless services are cool, but quite expensive.
- Add trigger rules when instantiating GCP services through airflow.
- Be sure what IAM permissions you need.



Next Steps

- Create Dim Tables and Fact table in GCP.
- Implement an NLP algorithm to classify movies with more accuracy.
- Add sensors to the DAG to identify failures.
- Use different approaches to test and improve performance.

Q&A

Thank you