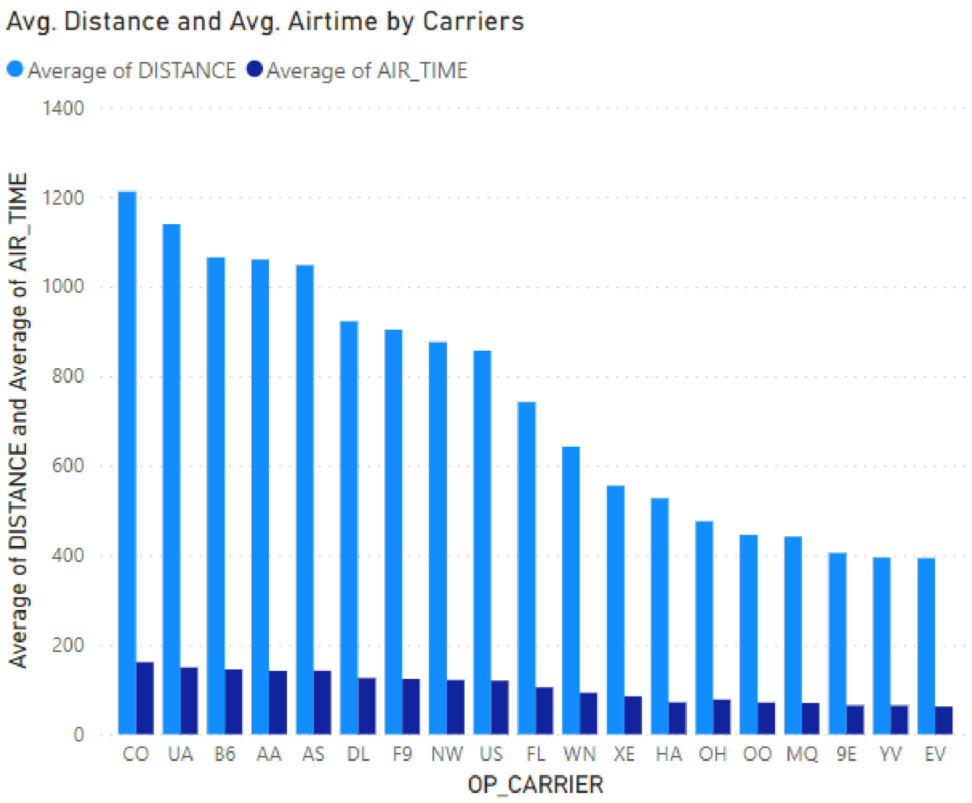
GROUP- 7

Members:
Ritesh Sengar
Shravan Honade
Nikhil Patil
Dhananjay Ghate

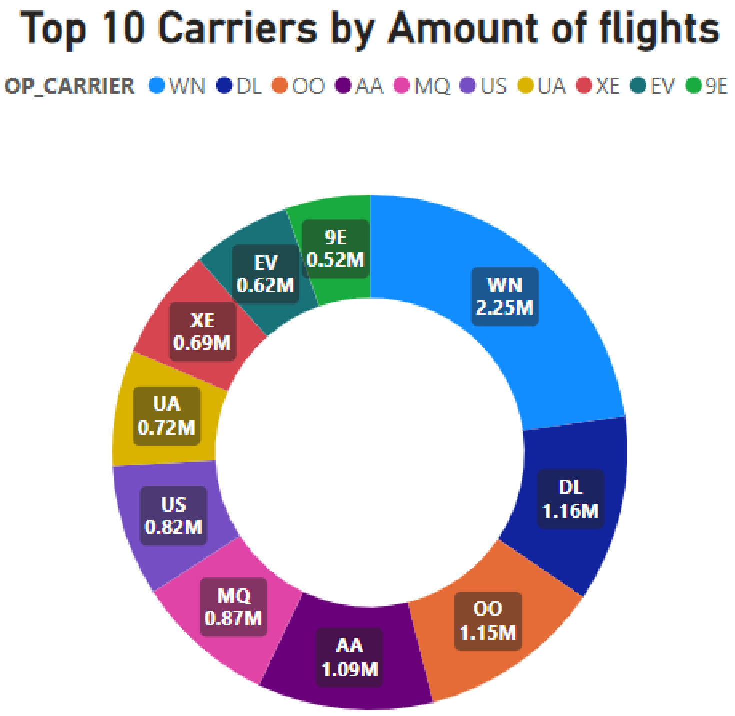
EDA



Here are some additional observations from the graph:

- Low-cost carriers (LCCs) such as Southwest Airlines (WN) and Spirit Airlines (NK) tend to have shorter average distances and shorter average airtimes than traditional carriers such as United Airlines (UA) and Delta Air Lines (DL). This is because LCCs typically operate shorter flights.

- Regional airlines such as Envoy Air (EV) and Piedmont Airlines (PI) tend to have shorter average distances and shorter average airtimes than major airlines. This is because regional airlines typically operate shorter flights to smaller airports.
- Charter airlines such as NetJets (XE) and Flexjet (HA) tend to have longer average distances and longer average airtimes than commercial airlines. This is because charter airlines typically operate flights to and from smaller airports and private airstrips.


Top 10 Carriers by Amount of flights

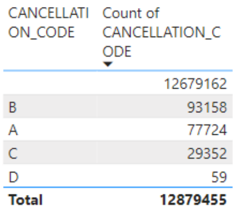
The pie chart shows the top 10 carriers by number of flights in 2017. The carriers are listed in order of decreasing number of flights, with American Airlines (AA) at the top and EV9E at the bottom.

The top 5 carriers by number of flights are:

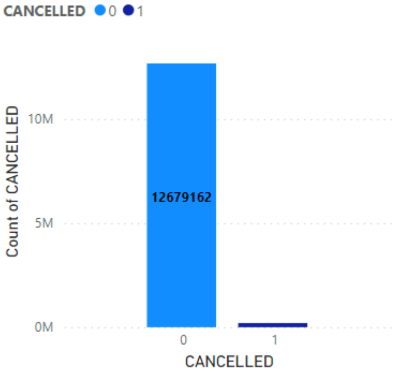
1. American Airlines (AA)
2. Delta Air Lines (DL)
3. Ryanair Group (RYR)
4. United Airlines (UA)
5. Southwest Airlines (WN)

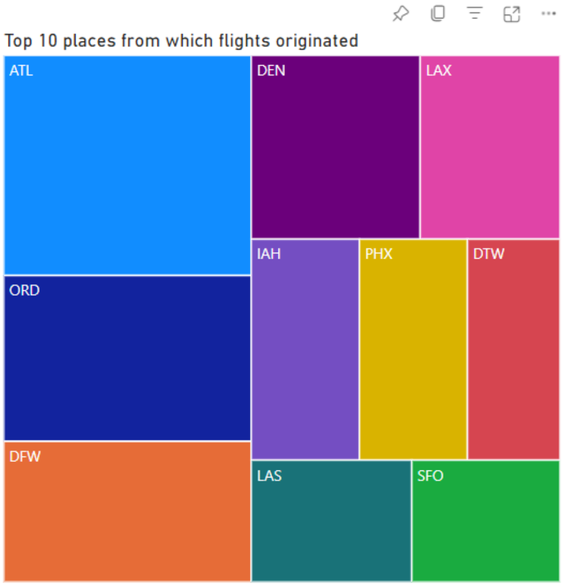
These carriers account for over 60% of all flights in 2017.

Total Flights cancelled from Top 10 originating Places:

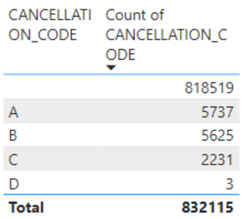


CANCELLATION_CODE	Count of CANCELLATION_CODE
	12679162
B	93158
A	77724
C	29352
D	59
Total	**12879455**

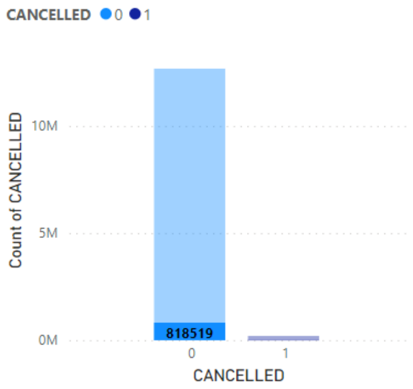


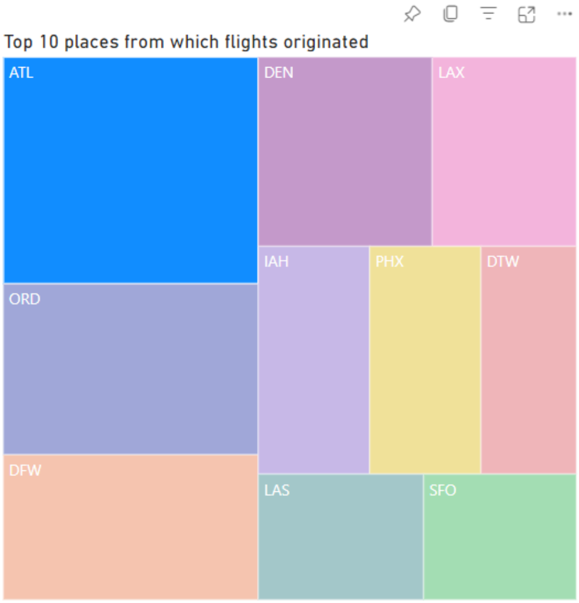


Maximum number of Flights cancelled from Originating Location:



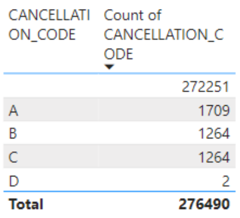
CANCELLATION_CODE	Count of CANCELLATION_CODE
	818519
A	5737
B	5625
C	2231
D	3
Total	**832115**



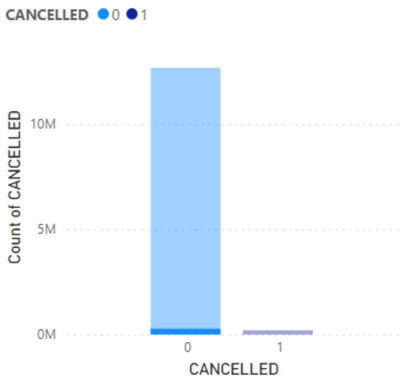


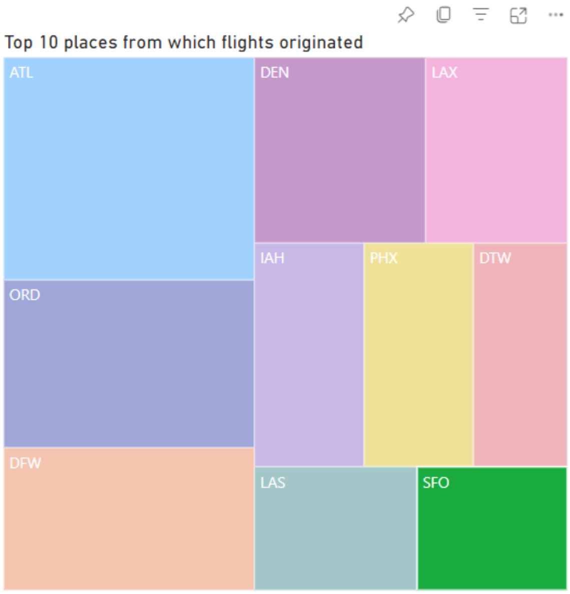
Top 10 places from which flights originated

Minimum number of flights cancelled from Originating Location:



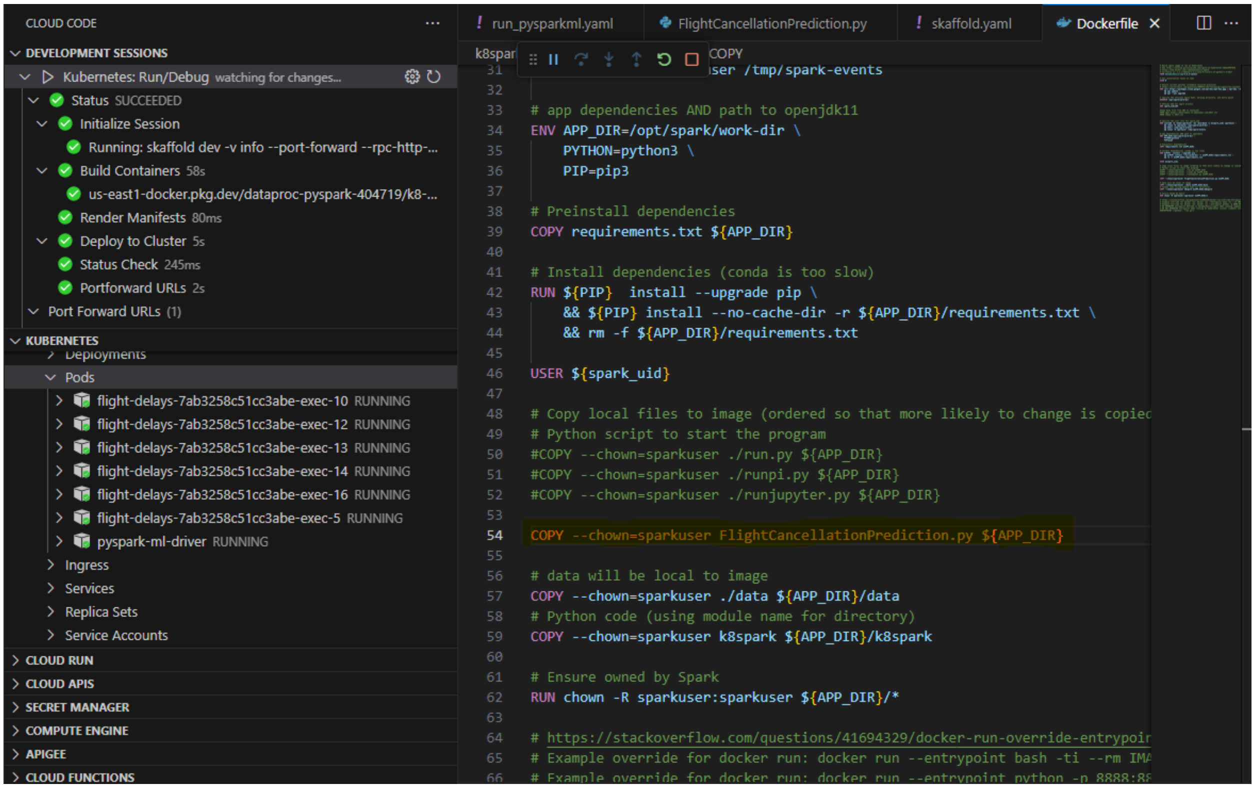
CANCELLATION_CODE	Count of CANCELLATION_CODE
	272251
A	1709
B	1264
C	1264
D	2
Total	**276490**





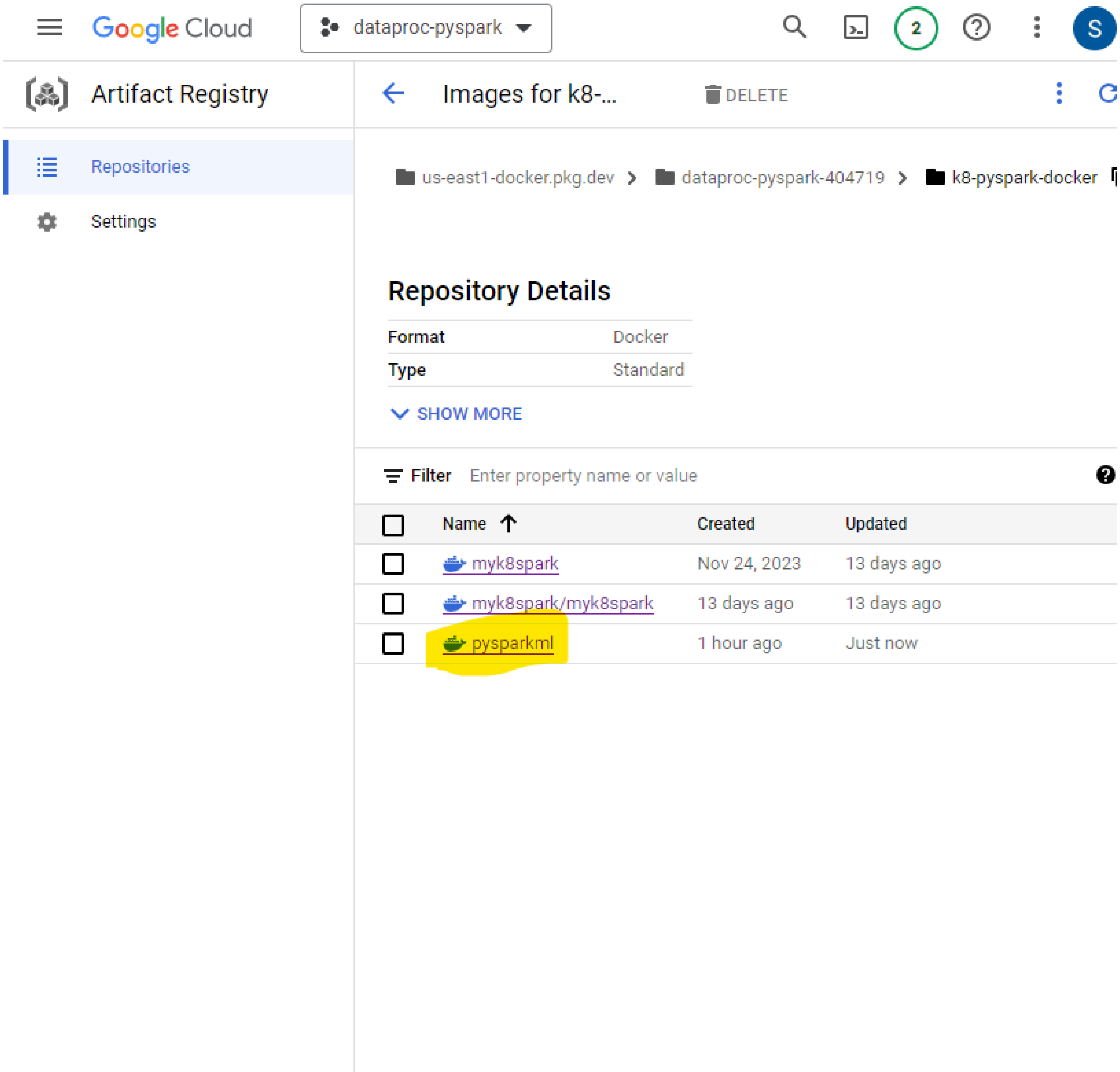
Top 10 places from which flights originated

PySpark Application deployment on Kubernetes:



```
CLOUD CODE
DEVELOPMENT SESSIONS
  Kubernetes: Run/Debug  watching for changes...
    Status  SUCCEEDED
      Initialize Session
        Running: skaffold dev -v info --port-forward --rpc-http-...
      Build Containers  58s
        us-east1-docker.pkg.dev/dataproc-pyspark-404719/k8-...
      Render Manifests  80ms
      Deploy to Cluster  5s
      Status Check  245ms
      Portforward URLs  2s
    Port Forward URLs  (1)
KUBERNETES
    Deployments
    Pods
      flight-delays-7ab3258c51cc3abe-exec-10  RUNNING
      flight-delays-7ab3258c51cc3abe-exec-12  RUNNING
      flight-delays-7ab3258c51cc3abe-exec-13  RUNNING
      flight-delays-7ab3258c51cc3abe-exec-14  RUNNING
      flight-delays-7ab3258c51cc3abe-exec-16  RUNNING
      flight-delays-7ab3258c51cc3abe-exec-5  RUNNING
      pyspark-ml-driver  RUNNING
    Ingress
    Services
    Replica Sets
    Service Accounts
CLOUD RUN
CLOUD APIS
SECRET MANAGER
COMPUTE ENGINE
APIGEE
CLOUD FUNCTIONS
```

```
ser /tmp/spark-events

# app dependencies AND path to openjdk11
ENV APP_DIR=/opt/spark/work-dir \
    PYTHON=python3 \
    PIP=pip3

# Preinstall dependencies
COPY requirements.txt ${APP_DIR}

# Install dependencies (conda is too slow)
RUN ${PIP}  install --upgrade pip \
    && ${PIP} install --no-cache-dir -r ${APP_DIR}/requirements.txt \
    && rm -f ${APP_DIR}/requirements.txt

USER ${spark_uid}

# Copy local files to image (ordered so that more likely to change is copiec
# Python script to start the program
#COPY --chown=sparkuser ./run.py ${APP_DIR}
#COPY --chown=sparkuser ./runpi.py ${APP_DIR}
#COPY --chown=sparkuser ./runjupyter.py ${APP_DIR}

COPY --chown=sparkuser FlightCancellationPrediction.py ${APP_DIR}

# data will be local to image
COPY --chown=sparkuser ./data ${APP_DIR}/data
# Python code (using module name for directory)
COPY --chown=sparkuser k8spark ${APP_DIR}/k8spark

# Ensure owned by Spark
RUN chown -R sparkuser:sparkuser ${APP_DIR}/*

# https://stackoverflow.com/questions/41694329/docker-run-override-entrypoir
# Example override for docker run: docker run --entrypoint bash -ti --rm IMA
# Example override for docker run: docker run --entrypoint python -p 8888:88
```



Images for k8-...

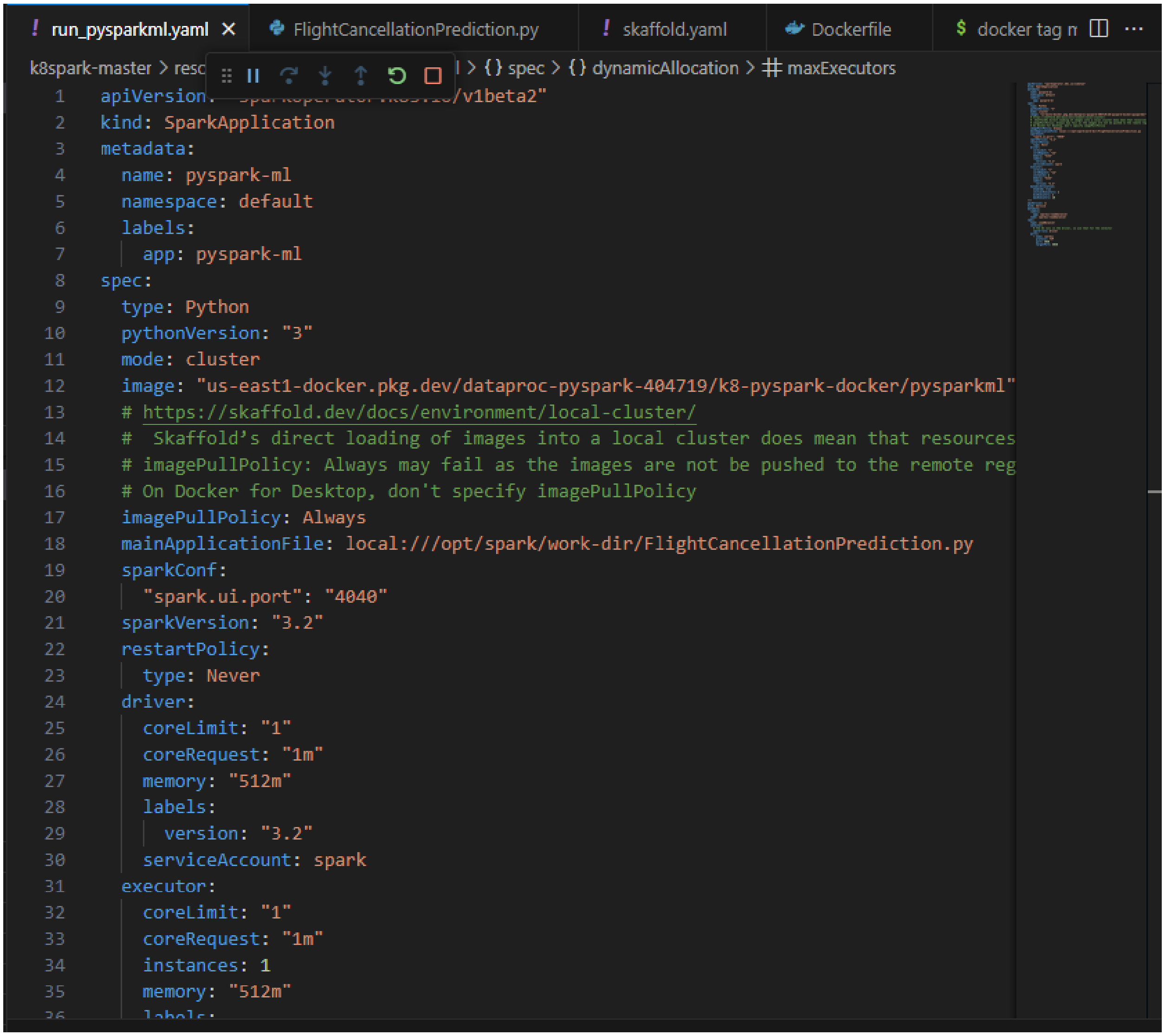
us-east1-docker.pkg.dev > dataproc-pyspark-404719 > k8-pyspark-docker

Repository Details

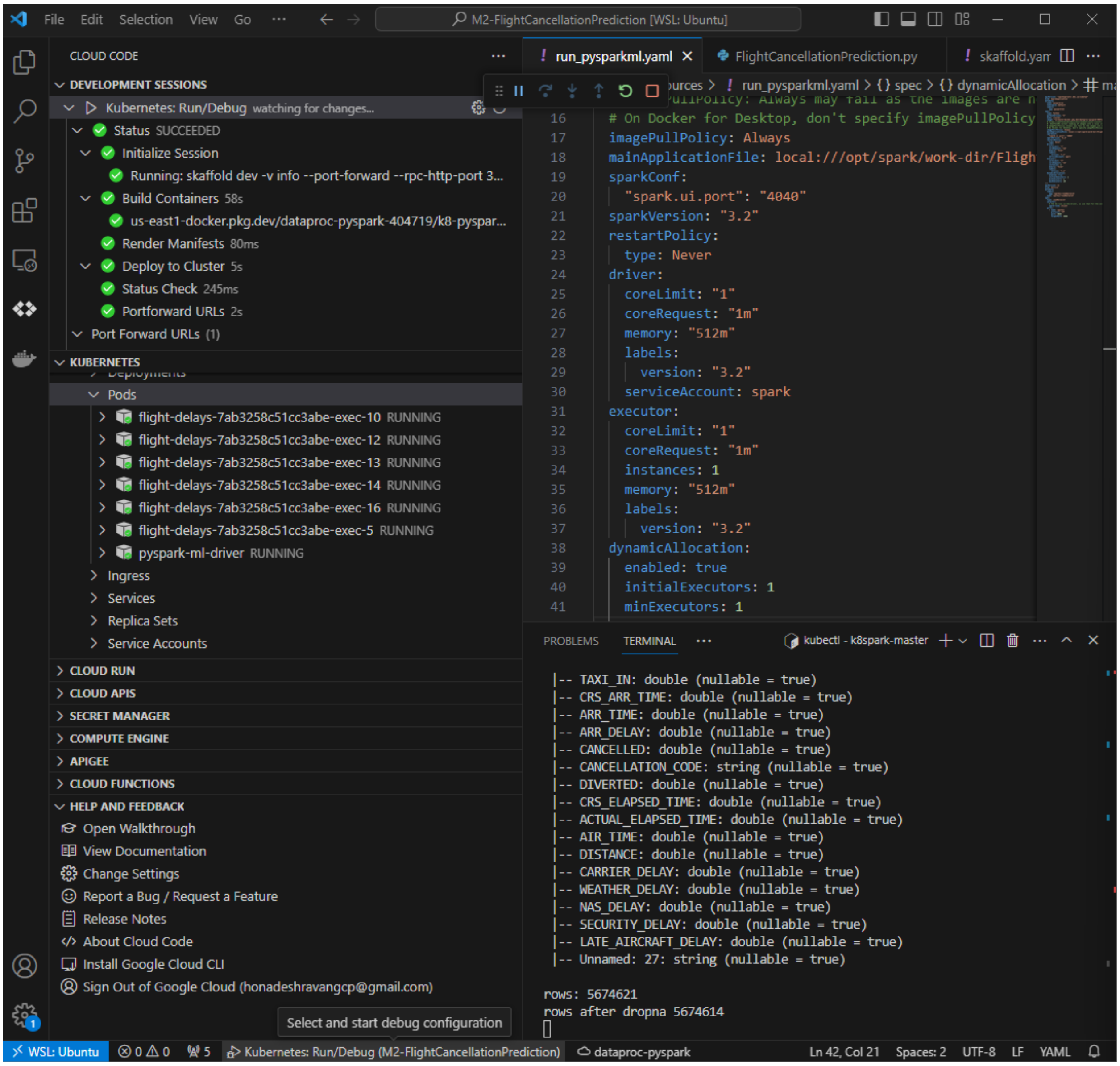
Format	Docker
Type	Standard

Enter property name or value

□	Name ↑	Created	Updated
□	myk8spark	Nov 24, 2023	13 days ago
□	myk8spark/myk8spark	13 days ago	13 days ago
□	pysparkml	1 hour ago	Just now


run_pysparkml.yaml
FlightCancellationPrediction.py
skaffold.yaml
Dockerfile
docker tag n
k8spark-master > reso
> { } spec > { } dynamicAllocation > # maxExecutors

```yaml
apiVersion: "sparkoperator.k8s.io/v1beta2"
kind: SparkApplication
metadata:
  name: pyspark-ml
  namespace: default
  labels:
    app: pyspark-ml
spec:
  type: Python
  pythonVersion: "3"
  mode: cluster
  image: "us-east1-docker.pkg.dev/dataproc-pyspark-404719/k8-pyspark-docker/pysparkml"
  # https://skaffold.dev/docs/environment/local-cluster/
  #  Skaffold's direct loading of images into a local cluster does mean that resources
  # imagePullPolicy: Always may fail as the images are not be pushed to the remote reg
  # On Docker for Desktop, don't specify imagePullPolicy
  imagePullPolicy: Always
  mainApplicationFile: local:///opt/spark/work-dir/FlightCancellationPrediction.py
  sparkConf:
    "spark.ui.port": "4040"
  sparkVersion: "3.2"
  restartPolicy:
    type: Never
  driver:
    coreLimit: "1"
    coreRequest: "1m"
    memory: "512m"
    labels:
      version: "3.2"
    serviceAccount: spark
  executor:
    coreLimit: "1"
    coreRequest: "1m"
    instances: 1
    memory: "512m"
    labels:
```


File
Edit
Selection
View
Go
M2-FlightCancellationPrediction [WSL: Ubuntu]
CLOUD CODE
run_pysparkml.yaml
FlightCancellationPrediction.py
skaffold.yam
urces > run_pysparkml.yaml > {} spec > {} dynamicAllocation > # ma

DEVELOPMENT SESSIONS

- Kubernetes: Run/Debug watching for changes...
  - Status SUCCEEDED
    - Initialize Session
      - Running: skaffold dev -v info --port-forward --rpc-http-port 3...
    - Build Containers 58s
      - us-east1-docker.pkg.dev/dataproc-pyspark-404719/k8-pyspar...
    - Render Manifests 80ms
    - Deploy to Cluster 5s
      - Status Check 245ms
      - Portforward URLs 2s
  - Port Forward URLs (1)

KUBERNETES

- Deployments
- Pods
  - flight-delays-7ab3258c51cc3abe-exec-10 RUNNING
  - flight-delays-7ab3258c51cc3abe-exec-12 RUNNING
  - flight-delays-7ab3258c51cc3abe-exec-13 RUNNING
  - flight-delays-7ab3258c51cc3abe-exec-14 RUNNING
  - flight-delays-7ab3258c51cc3abe-exec-16 RUNNING
  - flight-delays-7ab3258c51cc3abe-exec-5 RUNNING
  - pyspark-ml-driver RUNNING
- Ingress
- Services
- Replica Sets
- Service Accounts

```yaml
  ullPolicy: Always may fail as the images are n
  # On Docker for Desktop, don't specify imagePullPolicy
  imagePullPolicy: Always
  mainApplicationFile: local:///opt/spark/work-dir/Fligh
  sparkConf:
    "spark.ui.port": "4040"
  sparkVersion: "3.2"
  restartPolicy:
    type: Never
  driver:
    coreLimit: "1"
    coreRequest: "1m"
    memory: "512m"
    labels:
      version: "3.2"
    serviceAccount: spark
  executor:
    coreLimit: "1"
    coreRequest: "1m"
    instances: 1
    memory: "512m"
    labels:
      version: "3.2"
  dynamicAllocation:
    enabled: true
    initialExecutors: 1
    minExecutors: 1
```

```
 |-- TAXI_IN: double (nullable = true)
 |-- CRS_ARR_TIME: double (nullable = true)
 |-- ARR_TIME: double (nullable = true)
 |-- ARR_DELAY: double (nullable = true)
 |-- CANCELLED: double (nullable = true)
 |-- CANCELLATION_CODE: string (nullable = true)
 |-- DIVERTED: double (nullable = true)
 |-- CRS_ELAPSED_TIME: double (nullable = true)
 |-- ACTUAL_ELAPSED_TIME: double (nullable = true)
 |-- AIR_TIME: double (nullable = true)
 |-- DISTANCE: double (nullable = true)
 |-- CARRIER_DELAY: double (nullable = true)
 |-- WEATHER_DELAY: double (nullable = true)
 |-- NAS_DELAY: double (nullable = true)
 |-- SECURITY_DELAY: double (nullable = true)
 |-- LATE_AIRCRAFT_DELAY: double (nullable = true)
 |-- Unnamed: 27: string (nullable = true)

rows: 5674621
rows after dropna 5674614
```