**AWS Athena & QuickSight Homework**

This report summarizes the AWS Athena and Quicksight data analysis tasks completed using Athena SQL, notebook and quicksight. It includes step-by-step processes, Lake Formation security configuration, code snippets and analytical insights.

**Part 1. Setting up AWS Athena and Analyzing E-commerce Data**

**Step-by-step Process and Code Snippets**

1. **AWS Athena CloudFormation Template**

Create a CloudFormation template HW18-Athena-Part1.yaml that:

* Creates an S3 bucket.
* Creates a Glue Data Catalog Database for Athena SQL queries.
* Creates a Glue Crawler that crawls the database in the bucket.
* Creates an Athena WorkGroup with result output to the created S3 bucket.
* Registers the S3 path as a Lake Formation data location.
* Grants Lake Formation permissions:
  + Data location access to the Glue Crawler role
* Sets Lake Formation admins and default permissions. In this homework, admin is the root user.

After provisioning the AWS Lake Formation template, deploy it in AWS and wait until all the resources are correctly created.

A screenshot of a computer

AI-generated content may be incorrect.

1. **AWS Lake Formation Data Ingestion**

After the successful deployment of the CloudFormation template, direct to the S3 bucket, created a file named data, and upload the E-commerce database (link: <https://www.kaggle.com/datasets/carrie1/ecommerce-data>) to the created file.

A screenshot of a computer

AI-generated content may be incorrect.

Then, direct to the AWS Glue Console to start the crawler to get the dataset schema. Once the crawler successfully completes the task, the dataset schema can be found in the DataCatalog Tables in the AWS Glue Console.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Then, direct to AWS Lake Formation to check the S3 bucket is successfully registered in Lake Formation.

A screenshot of a computer

AI-generated content may be incorrect.

1. **Athena Data Analysis**

The following data analysis tasks are completed in Athena using Athena SQL queries.

1. Find the top 10 most purchased products.

**Code Snippets and Results:**

select stockcode as ProductsNumber, count (\*) as product\_count

from "AwsDataCatalog"."hw18-ecommerce"."data"

where quantity > 0

group by stockcode

order by product\_count desc

limit 10

A screenshot of a computer

AI-generated content may be incorrect.

It can be concluded from the result that product 85123A, 85099B, 22423, 47566, 20725, 84879, 22197, 22720, 21212, 20727 are the top 10 most purchased products.

1. Calculate the Average Spend per Customer

**Code Snippets and Results:**

select customerid, sum(quantity \* unitprice) / count(distinct invoiceno) as average\_spend

from "AwsDataCatalog"."hw18-ecommerce"."data"

where customerid is not null and quantity > 0

group by customerid

order by average\_spend desc

A screenshot of a computer

AI-generated content may be incorrect.

It can be concluded that the average spending per customer ranges from 4313 to 77183.6.

1. Calculate the Total Revenue for Each Month

**Code Snippets and Results:**

select

date\_format(

coalesce(

try(date\_parse(trim(InvoiceDate), '%c/%e/%Y %k:%i:%s')),

try(date\_parse(trim(InvoiceDate), '%c/%e/%Y %k:%i'))

),

'%Y-%m'

) as month,

sum(quantity \* unitprice) as revenue

from "AwsDataCatalog"."hw18-ecommerce"."data"

where coalesce(

try(date\_parse(trim(InvoiceDate), '%c/%e/%Y %k:%i:%s')),

try(date\_parse(trim(InvoiceDate), '%c/%e/%Y %k:%i'))

) is not null

group by 1

order by 1

A screenshot of a computer

AI-generated content may be incorrect.

It can be concluded that for each month, the revenue ranges from 431427 to 744310.

1. Identify the Day with the Highest Revenue

**Code Snippets and Results:**

select

date\_format(

coalesce(

try(date\_parse(trim(InvoiceDate), '%c/%e/%Y %k:%i:%s')),

try(date\_parse(trim(InvoiceDate), '%c/%e/%Y %k:%i'))

),

'%Y-%m-%d'

) as day,

sum(quantity \* unitprice) as revenue

from "AwsDataCatalog"."hw18-ecommerce"."data"

where coalesce(

try(date\_parse(trim(InvoiceDate), '%c/%e/%Y %k:%i:%s')),

try(date\_parse(trim(InvoiceDate), '%c/%e/%Y %k:%i'))

) is not null

group by 1

order by revenue desc

limit 10

A screenshot of a computer

AI-generated content may be incorrect.

It can be concluded that the highest revenue occurred on 14th November 2011, with revenue 111893.

**Part 2. Analyzing Flight Delays with Athena and Spark**

**Step-by-step Process and Code Snippets**

1. **Data Ingestion**

In this part, we still use the CloudFormation template in Homework 1. Firstly, the dataset in Part 1 is deleted and the Flight Delays Data is uploaded to the S3 bucket. The link to the dataset: https://www.kaggle.com/datasets/sriharshaeedala/airline-delay

A screenshot of a computer

AI-generated content may be incorrect.

After successfully uploading the Flight Delays dataset, recrawl the data and the new table is created in AWS Glue.

1. **Athena Notebook Data Analysis**

Create an Athena Notebook in AWS Athena console called Athena\_Part2 to conduct the data analysis.

The following data analysis tasks are completed in Athena Notebook.

1. View the DataFrame's schema and the first ten rows.

**Code Snippets and Results:**

from pyspark.sql import functions as f

s3\_input = "s3://hw18-part1/data/Airline\_Delay\_Cause.csv"

df = spark.read.option("header", "true").option("inferSchema", "true").csv(s3\_input)

df.printSchema()

df.show(10, truncate=False)

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

The dataset contains 21 columns. Each record is a (year, month, carrier, airport) tuple. It includes identifiers, traffic totals (arrivals/cancellations/diversions), delay counts by cause (≥15min), and delay minutes by cause as well as overall

1. Describe the statistical properties of the columns.

**Code Snippets and Results:**

# describe the statistical properties

desc\_df = df.describe()

desc\_df.show()

A screenshot of a computer

AI-generated content may be incorrect.

The dataset contains 171,666 records. Categorical fields (carrier/ carrier\_name/ airport/ airport\_name) are present for all rows (e.g., carriers range from 9E to YX, airports from ABE to YUM). For numeric fields, arr\_flights has 171,426 non-null values with a mean of 362.53, stddev 992.89, min 1, max 21,977. arr\_del15 (flights delayed ≥15 minutes) has 171,223 non-null values with a mean of 66.43, stddev 179.54, min 0, max 4,176. Delay-by-cause count columns (carrier\_ct, weather\_ct, nas\_ct, security\_ct, late\_aircraft\_ct) and the corresponding delay-minutes columns exhibit similar sparsity and long tails.

1. Identify the top 5 airlines with the highest average delays.

**Code Snippets and Results:**

# identify the airlines with top 5 delays

delays\_df = (df.groupBy("carrier\_name")

.agg(f.avg("arr\_delay").alias("Average\_Delays"))

.orderBy(f.desc(f.col("Average\_Delays"))).limit(5))

delays\_df.show()

A close up of a number

AI-generated content may be incorrect.

It can be concluded from the result table that the top 5 airlines with the highest average delays are Southwest Airlines, American Airlines, United Airlines, JetBlue Airways and Spirit Airlines.

1. Discover the 3 airports with the most cancellations

**Code Snippets and Results:**

# discover the top three airports with the most cancellations

airports\_df = (df.groupBy("airport","airport\_name")

.agg(f.sum("arr\_cancelled").alias("Cancellation\_Count"))

.orderBy(f.desc(f.col("Cancellation\_Count")))

.limit(3)

)

airports\_df.show()

A close up of a document

AI-generated content may be incorrect.

It can be concluded from the result table that ORD, DFW, and LGA are the top 3 airport with most of the cancellations, among which ORD has the maximum cancellation of 80821.

1. Determine the most common reason for flight delays

**Code Snippets and Results:**

# determine the most common reason for flight delays

reason\_df = (df.select(

f.sum("carrier\_delay").alias("Carrier"),

f.sum("weather\_delay").alias("Weather"),

f.sum("nas\_delay").alias("NAS"),

f.sum("security\_delay").alias("Security"),

f.sum("late\_aircraft\_delay").alias("Late\_Aircraft"))

)

reason\_long = (

reason\_df.selectExpr(

"stack(5,"

" 'Carrier', Carrier,"

" 'Weather', Weather,"

" 'NAS', NAS,"

" 'Security', Security,"

" 'Late\_Aircraft', Late\_Aircraft"

") as (reason, total\_delay)"

)

.orderBy(f.desc("total\_delay"))

)

reason\_long.show()

reason\_long.limit(1).show()

A screenshot of a computer program

AI-generated content may be incorrect.

It can be concluded from the table that the most common reason for delay is due to late aircraft. This implies that plane arrival has a huge impact on flight delays.

**Part 3. In-depth Data Analysis of World Development Indicators with QuickSight**

**Step-by-step Process and Code Snippets**

1. **Data Ingestion and ETL Pipeline**

Create an S3 bucket named hw18-part3 and create a file data in the bucket, then download the World Development Indicator (WDI) dataset and upload the data (WDICSV) to the bucket.

The link to the dataset: <https://www.kaggle.com/datasets/theworldbank/world-development-indicators>

Then, set up a yaml file named: HW18-Athena-Part3\_Role to provision necessary IAM role for the dataset to grant access to perform ETL pipeline in AWS Glue. Deploy the yaml file in the AWS CloudFormation and wait for the successful creation of the necessary resources.

Then, perform the ETL pipeline in AWS Glue. In this scenario, we will do the following to clean and transform the data.

**Keep the Key Indicators and Years**

We would like to keep the key indicators (GDP, population growth, and health expenditure) and the years we want to analyze. Therefore, we extract the key indicators’ columns and years’ columns.

A computer screen with text

AI-generated content may be incorrect.

**Change the Chart Format**

To support deeper analysis in QuickSight, we will reshape the dataset into a tidy long format with columns: **Country Name**, **Country Code,** **Indicator Name**, **Indicator Code,** **Year**, and **Value**.

A computer code on a black background

AI-generated content may be incorrect.

**Filter our Null Value and Duplicate Records**

To ensure data quality in our analysis, we dropped rows with nulls and removed duplicates based on Country, Indicator, Year.

A screen shot of a computer

AI-generated content may be incorrect.

After the ETL pipeline is completed, write out the transformed data to the designated S3 bucket file.

1. **QuickSight Analysis**

To begin with the data analysis, the transformed dataset should be uploaded to QuickSight. Therefore, a Json file named HW18-QuickSight-Manifest containing the file locations in the S3 bucket and upload settings is provisioned. Upload the Json file in the QuickSight to load the data to begin with the analysis.

* 1. **Descriptive Analysis**

To characterize the distribution of key indicators, we use bar charts with indicator names on the x-axis and the mean value (averaged across countries within the selected years) on the y-axis, among which GDP is plotted on a logarithmic scale, the rest indicators distribution results are plotted on a normal scale.

**GDP Distribution Analysis:**

**A graph of colorful lines

AI-generated content may be incorrect.**

It can be concluded from the chart that only a small set of economies exceeds USD 1 trillion, while most countries fall between USD 0.1–1 trillion. This pattern indicates a strong right-skew and a high concentration of global output in a few large economies.

**Population Growth Distribution Analysis:**

A graph of different colored lines

AI-generated content may be incorrect.

From the Population Growth Distribution Analysis, most countries record an average annual growth rate of about 2–3% over the selected years. A small number of economies—especially some Gulf states such as Qatar and the UAE—show higher rates (around 6–7%), often driven by sustained labor immigration and rapid urban expansion rather than natural increase alone. Overall, the bulk of global population growth occurs in lower- and middle-income countries, reflecting younger age structures, higher fertility, and improving child survival. These patterns imply rising demand for infrastructure, education, and health services in those regions.

**Health Expenditure Distribution Analysis:**

A graph of different colored lines

AI-generated content may be incorrect.

From the Health Expenditure (% of GDP) distribution results, most countries allocate around 9% of GDP to healthcare in the selected years. A smaller group—primarily high-income economies and some small island states—spend above 10–12%. For high-income countries this typically reflects broader coverage, older populations, and higher input prices; for small islands, scale constraints and reliance on imported medical services can make health spending a larger portion of GDP.

* 1. **GDP Comparison in Different Countries**

In this scenario, we use bar charts with indicator names on the x-axis and the total value (averaged across countries within the selected years) on the y-axis, among which GDP is plotted on a logarithmic scale.

**A graph of value

AI-generated content may be incorrect.**

From the GDP result, most of the countries’ cumulative GDP over the selected years falls between USD 10T and 100T. A small group of advanced economies (e.g., the United States, China, Japan) exceed 100 trillion. Several countries are below 10T, which mainly reflects a smaller economic base and population, rather than poverty alone.

* 1. **Correlation Analysis**

To conduct the correlation analysis, the chart is transformed into 4 columns: Country, Current Health Expenditure, GDP, and Population Growth and stored in an excel sheet called: Countries with Key Indicator Values.xlsx.

To perform the correlation analysis between key indicators and GDP, we use scatter plot with cumulative GDP on the x-axis on logarithmic scale and the other indicator on a normal scale on the y-axis.

**GDP Correlation Analysis with Population Growth:**

A graph with many colored dots

AI-generated content may be incorrect.

From the chart, we can observe a clear pattern between the GDP and Population Growth. Countries with a GDP value larger than 10 trillion tend to have lower population growth, most of which are less than 2% annual increase. In contrast, most low GDP or middle GDP countries have a population annual growth rate of more than 2 percent. However, there are some outliers, especially gulf countries, which have high population annual growth but middle level GDP. This may suggest that Gulf countries have high immigration. These patterns suggest that future demand growth in headcount is concentrated in emerging markets, while advanced economies will rely more on productivity gains or migration to sustain growth.

**GDP Correlation Analysis with Health Expenditure:**

A graph with many colored dots

AI-generated content may be incorrect.

From the chart, we can observe a clear moderate positive association between economic size and current health expenditure. Countries with a GDP value larger than 1 trillion tend to have higher population growth, most of which are more than 3%. In contrast, most low GDP or middle GDP countries have a health expenditure percent between 3 to 9 percent. However, there are some outliers, especially for the island countries like Nauru and Marshall Islands. Those countries have low GDP but spend more than 12 percent on health care services. For these island countries, fixed system costs, reliance on imported medical goods/services, and small populations can push the ratio up even when absolute spending is modest.

* 1. **Time-Series Analysis**

To perform the time-series analysis, we select line chart with selected years on the x-axis and GDP for each year on logarithmic scale on the y-axis. For this analysis, we select the GDP data from the past two decades (2005 to 2024) in five countries: United States, United Kingdom, China, France, and Brazil.

A graph of different colored lines

AI-generated content may be incorrect.

From the line chart, we can observe the following:

For the United States, it stays the largest economy in the selected countries during this period. Although it experienced a small decrease in 2008 and 2020, the GDP is still going upwards in the past two decades.

For China, the GDP rises sharply in the past two decades, especially from 2005 to 2021. However, from 2021-2024, the GDP increase has slowed.

For UK and France, both exhibit broadly flat to mildly rising trajectories with small cyclical movements—softness after 2008–09, partial recovery in the mid-2010s, a dip in 2020, and only modest changes thereafter.

For Brazil, the GDP rises sharply between 2005 to 2011, then it persistently declines until 2020. Although the GDP rises after 2020, it is still below the peak in 2010s.

From the GDP situation in the five selected countries, the following patterns can be concluded:

United States remains the country with the highest GDP in the previous two decades.

France and UK’s GDP experienced cyclical changes in the past two decades.

Brazil and China both experienced sharp GDP rises in the past two decades. However, China remains the trend until 2020, Brazil experienced prolonged decline in the mid-2010s and still cannot exceed the peak GDP.

From the result chart, two synchronized downturns abnormalities can be noticed, signifying two global shocks that have a negative effect on most of the countries’ GDP. In 2008–09 (the global financial crisis) and in 2020 (the COVID-19 recession), the US, UK, France, and Brazil all show year-over-year declines in GDP. China is the exception: it slows markedly in both episodes but does not show the same depth of contraction as the others.