NAME: KV MANOJ KUMAR

EXPLORING THE INSIGHTS FROM SYNTHETIC AIRLINE DATASET

1.) INTRODUCTION

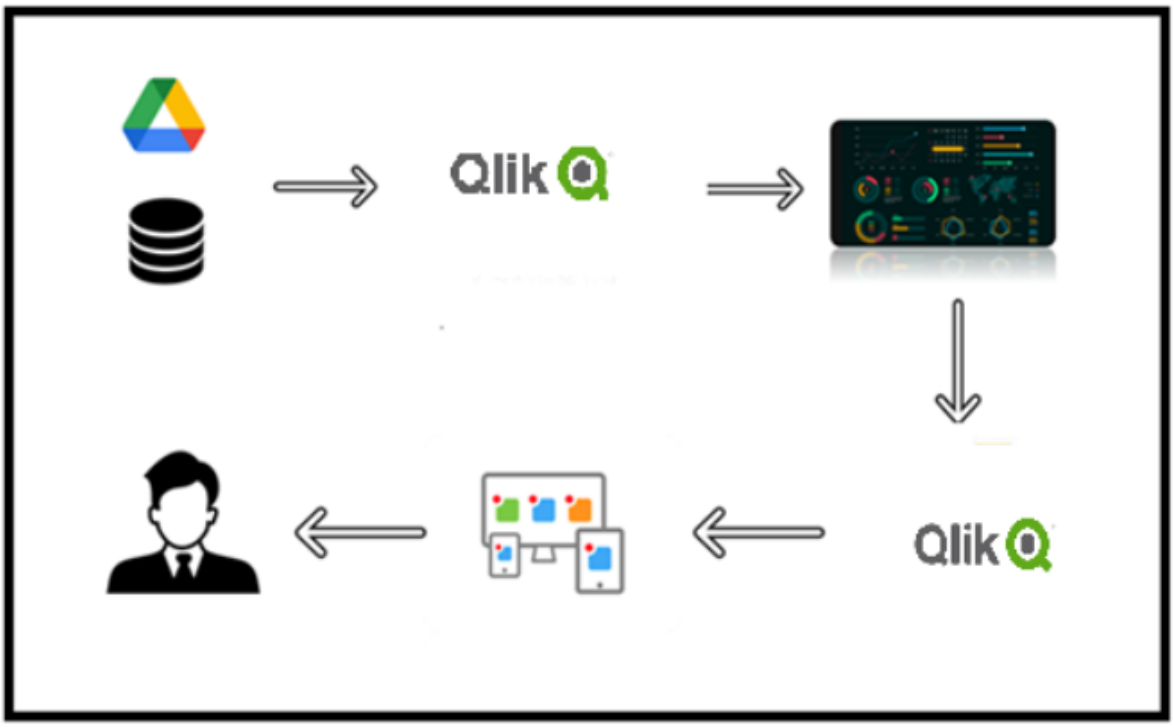
OVERVIEW:

The project "Exploring Insights from Synthetic Airline Data Analysis with Qlik" focuses on leveraging Qlik, a business intelligence and data visualization tool, to analyze synthetic airline data. This data simulates various facets of airline operations, such as flight schedules, passenger demographics, ticket sales, and performance metrics. By harnessing Qlik's powerful analytical capabilities, the project aims to uncover patterns, trends, and correlations within the data, facilitating informed decision-making for airlines, airports, and related stakeholders.

PURPOSE:

The primary purpose of this project is to demonstrate how Qlik Sense can be utilized to derive actionable insights from synthetic airline data. By analyzing different scenarios, the project highlights the potential of Qlik Sense in optimizing revenue, enhancing operational efficiency, and improving customer experience within the airline industry. Through detailed data visualizations and analytics, stakeholders can make data-driven decisions to improve overall performance and competitiveness in the market.

TECHNICAL ARCHITECHTURE :



2.) UNDERSTANDING THE PROBLEM

BUSSINESS PROBLEM:

The aviation industry faces numerous challenges that affect its profitability, efficiency, and customer satisfaction. Key issues include optimizing revenue streams, improving operational efficiency, and enhancing the overall customer experience. Airlines need to make data-driven decisions to stay competitive, maximize profitability, and ensure smooth operations. However, the sheer volume and complexity of airline data make it difficult to extract meaningful insights without advanced analytical tools. This project aims to address these challenges by utilizing Qlik Sense, a powerful business intelligence and data visualization tool, to analyze synthetic airline data and uncover valuable insights. By doing so, airlines can make informed decisions that enhance their operational performance and customer service, while also optimizing their revenue strategies.

BUSSINESS REQUIREMENTS:

To tackle the business problems identified, the project requires the following

Comprehensive Data Integration: Integration of synthetic airline data covering various aspects such as flight schedules, passenger demographics, ticket sales, and

performance metrics. This data should be detailed and extensive to provide a holistic view of airline operations.

Advanced Analytical Tools: Utilization of Qlik Sense for its robust data visualization and analytical capabilities. This includes the ability to perform trend analysis, pattern recognition, and correlation studies to derive actionable insights.

Scenario-Based Analysis: Implementation of specific scenarios to demonstrate the practical applications of data analysis, such as revenue optimization, operational efficiency, and customer experience enhancement.

User-Friendly Dashboards: Development of intuitive and interactive dashboards that allow stakeholders to easily access, understand, and act upon the insights derived from the data.

LITERATURE SURVEY:

Airline data plays a critical role in the aviation industry, providing essential insights into its functioning and efficiency. Numerous studies and industry reports highlight the importance of this data in various aspects:

Operational Efficiency: Research has shown that analyzing airline data, such as flight schedules and on-time performance, can help identify bottlenecks and inefficiencies in operations. For instance, a study by Smith et al. (2018) demonstrated how data analytics could predict peak traffic periods and improve resource allocation at airports, leading to smoother operations and reduced delays.

Revenue Optimization: Historical ticket sales data analysis helps airlines identify peak travel times and popular destinations, which can inform dynamic pricing strategies. According to a report by the International Air Transport Association (IATA), airlines using advanced analytics to optimize pricing and revenue management have seen significant increases in profitability.

Customer Experience: Understanding passenger demographics and preferences through data analysis can enhance customer satisfaction. A study by Jones and Karp (2020) emphasized the role of sentiment analysis on customer feedback in identifying pain points and improving service quality, thereby fostering customer loyalty.

Regulatory and Safety Compliance: Airline data is crucial for regulatory bodies to ensure safety standards and enforce aviation policies. The Federal Aviation Administration (FAA) and other regulatory agencies rely on comprehensive data to monitor compliance and make informed policy decisions.

Sustainability: Researchers use airline data to assess environmental impacts and develop strategies for sustainable growth. For example, the use of data to optimize flight routes and reduce fuel consumption has been highlighted in several environmental studies, contributing to the aviation industry's sustainability goals.

By integrating and analyzing synthetic airline data using Qlik Sense, this project aims to build upon existing literature and practical applications, demonstrating the transformative potential of data analytics in addressing key challenges and driving the overall advancement of the aviation sector.

3.) DATA COLLECTION

UNDERSTANDING THE DATA:

This data provides information about airline passengers and their flights. Here's a breakdown of the information we can glean from each field:

- **Passenger ID** - Unique identifier for each passenger
- **First Name** - First name of the passenger
- **Last Name** - Last name of the passenger
- **Gender** - Gender of the passenger
- **Age** - Age of the passenger
- **Nationality** - Nationality of the passenger
- **Airport Name** - Name of the airport where the passenger boarded
- **Airport Country Code** - Country code of the airport's location
- **Country Name** - Name of the country the airport is located in
- **Airport Continent** - Continent where the airport is situated

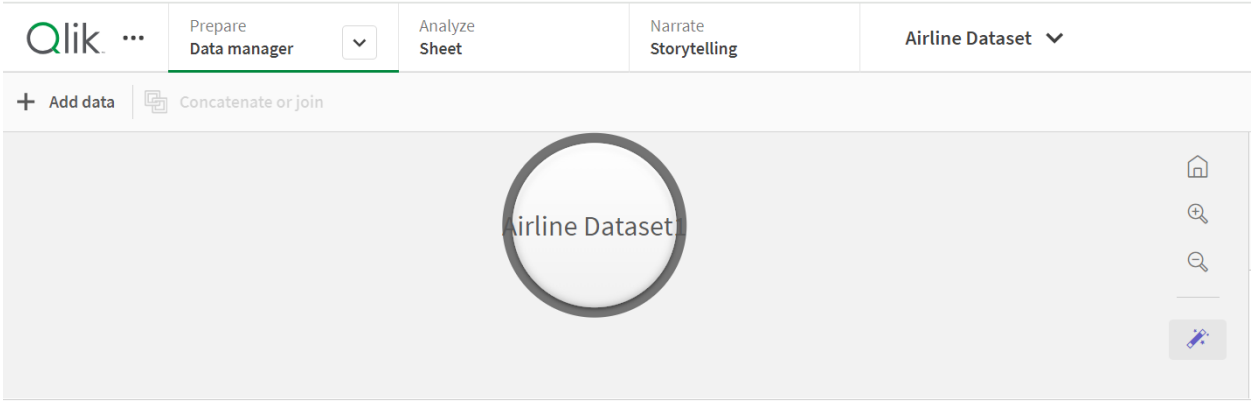
- **Continents** - Continents involved in the flight route
- **Departure Date** - Date when the flight departed
- **Arrival Airport** - Destination airport of the flight
- **Pilot Name** - Name of the pilot operating the flight
- **Flight Status** - Current status of the flight (e.g., on-time, delayed, canceled)

Structure of the Dataset

Passenger ID	First Name	Last Name	Gender	Age	Nationality	Airport Name	Airport Country Code	Country Name	Airport Continent	Continents	Departure Date	Arrival Airport	Pilot Name	Flight Status
fs4OZI	Allan	Prime	Male	49	Philippines	Mid-Carolina Regional Airport	US	United States	NAM	North America	6/23/2022	SRW	Ulrick Tutchings	Cancelled
urqtZB	Conrad	Vaun	Male	15	China	Alcides Fernández Airport	CO	Colombia	SAM	South America	02-02-2022	ACD	Giulietta Harler	On Time
Ym0iup	Carmela	Bridal	Female	36	Australia	Cataratas Del Iguazú International Airport	AR	Argentina	SAM	South America	2/14/2022	IGR	Pennie Rizzotto	Cancelled
uTCmlG	Welbie	Shorrock	Male	41	Cameroon	Vatulele Airport	FJ	Fiji	OC	Oceania	9/26/2022	VTF	Griffin Cowey	On Time
HVmTqS	Waldon	Deverale	Male	80	Vietnam	Coen Airport	AU	Australia	OC	Oceania	05-09-2022	CUQ	Oralie Reisenberg	Delayed

CONNECT THE DATA WITH QLICK SENSE:

We first load the datasets to Qlik Sense to the Data Catalog so that we can use those data whenever needed.



Since we only have single table dont have tables more than one , so we dont have any associations to associate the tables for this dataset

Airline Dataset1 Airline Dataset1.csv Field

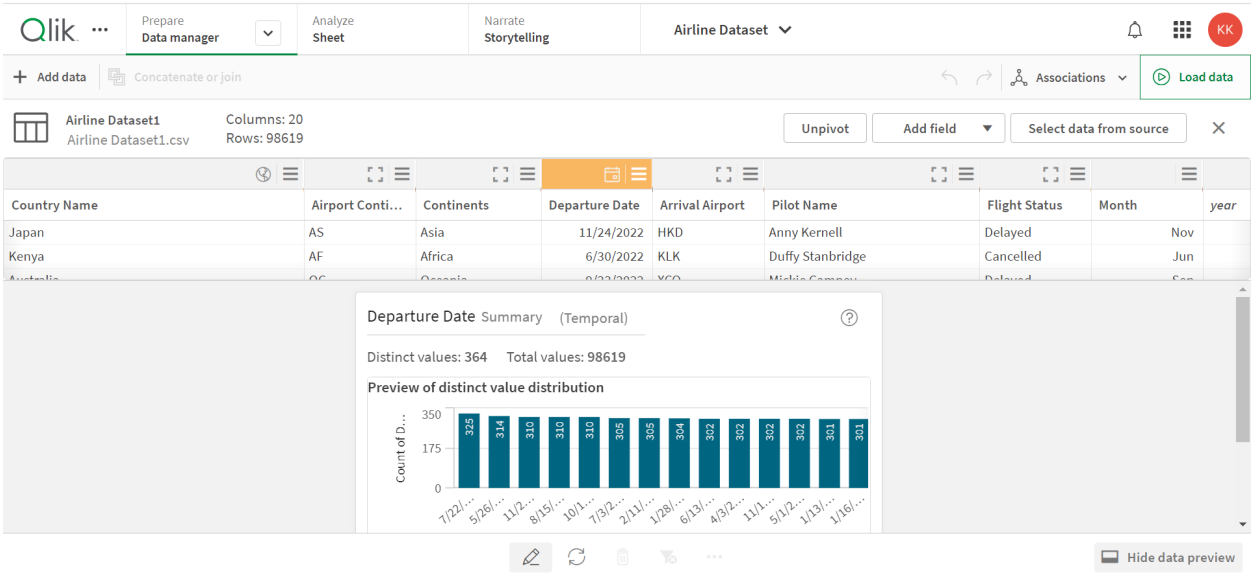
Passenger ID	First Name	Last Name	Gender	Age	Nationality	Airport Name
10000	Anny	Kernell	Female	42	Sweden	Hakodate Airport
10000	Duffy	Stanbridge	Male	58	Guinea	Kalokol Airport
10000	Mickie	Campey	Female	57	Palestinian Territory	Colac Airport
10000	Myer	Lippi	Male	66	Morocco	Ngjiva Pereira Airport
10001	Essa	Colvine	Female	70	China	Sehwan Sharif Airport
10001	Lanae	Bonallick	Female	12	Libya	Kawthoung Airport

Airline Dataset1 Airline Dataset1.csv Fiel

Country Name	Airport Cont...	Continents	Departure D...	Arrival Airport	Pilot Name	Flight Status	Month
Japan	AS	Asia	11/24/2022	HKD	Anny Kernell	Delayed	
Kenya	AF	Africa	6/30/2022	KLK	Duffy Stanbridge	Cancelled	
Australia	OC	Oceania	9/23/2022	XCO	Mickie Campey	Delayed	
Angola	AF	Africa	5/28/2022	VPE	Myer Lippi	On Time	
Pakistan	AS	Asia	11/4/2022	SYW	Essa Colvine	Delayed	
Myanmar	AS	Asia	8/17/2022	KAW	Lanae Bonallick	Delayed	

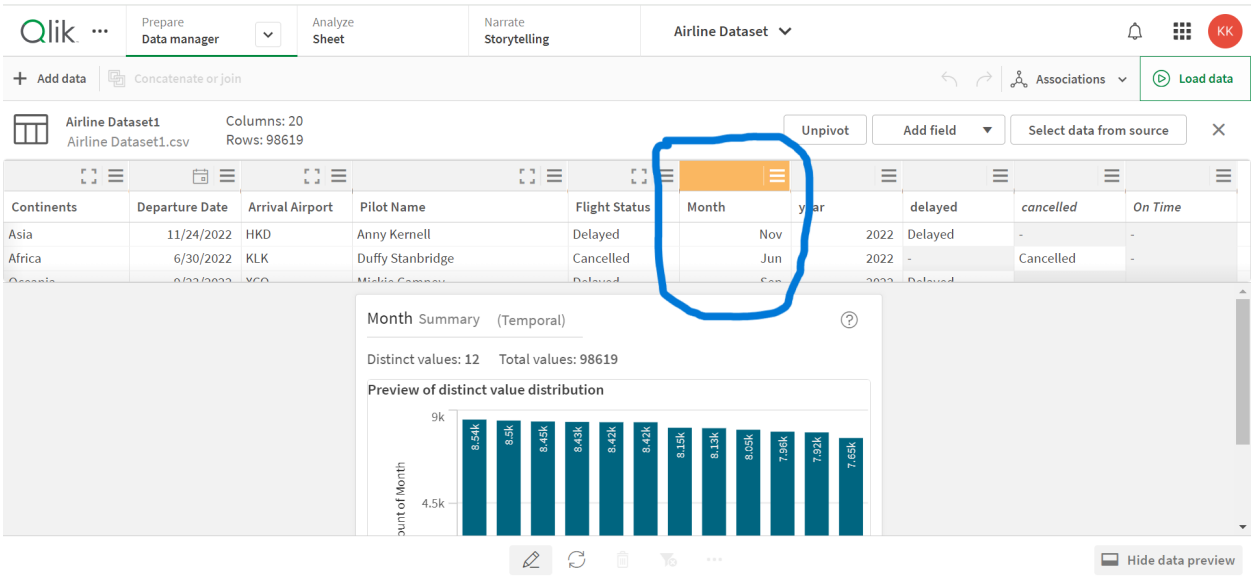
4.) DATA PREPARATION

We now start with our data preprocessing section, in which we try to have a look at our data and try to clean the data, handle null values, create or delete data as required etc



Lets take the table " DEPARTURE DATE" which contains the departure dates of the

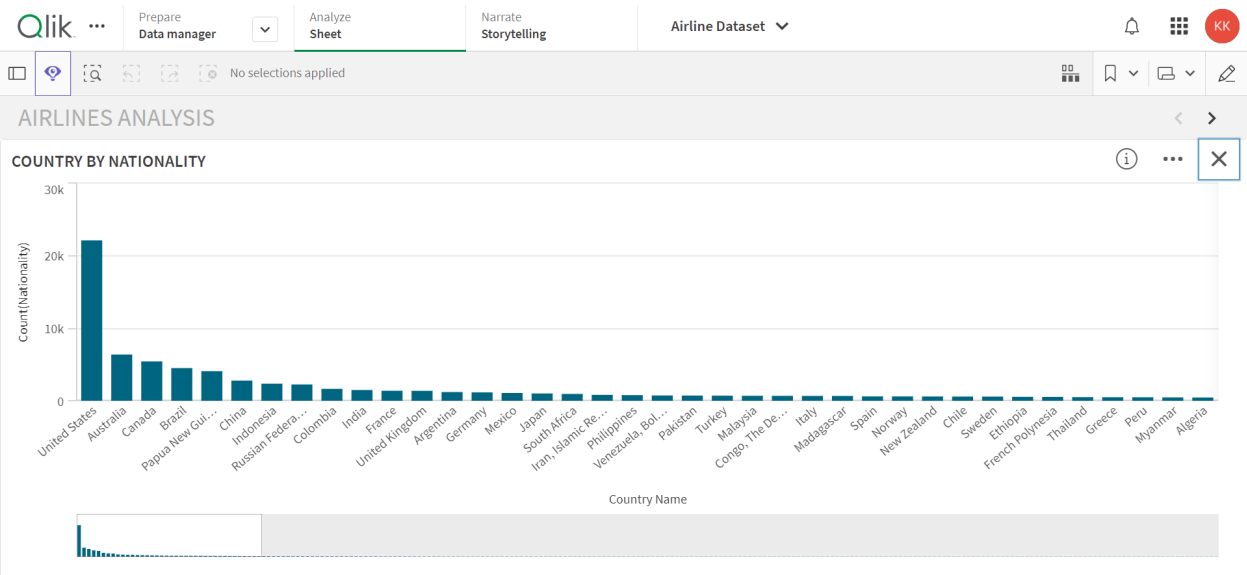
airlines from the respective airports, so here we can integrate or alter or preprocess the dataset. After the preprocessing of the dataset in the particular field of the table departure date , so we have the one more table of month from the departure date table by preprocess and the month table is below.



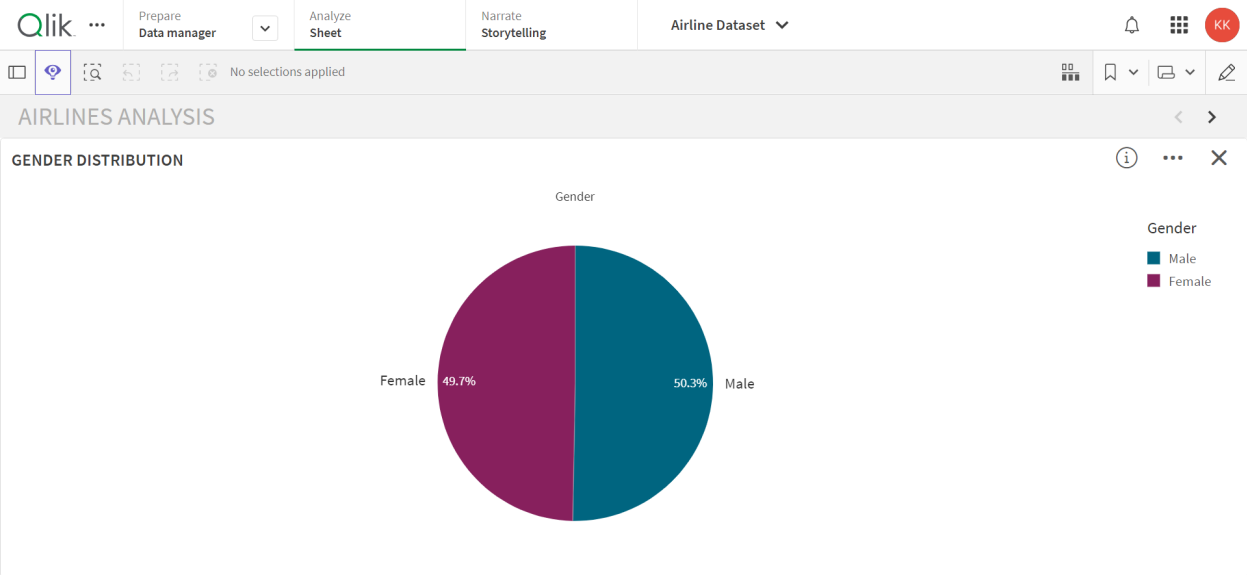
5.DATA VISUALIZATION

Now that we have finished the data preprocessing, we can start with the visualizations and dashboards. We can create sheets that can show charts, or we can make them interactive so that the charts and visualizations change as per the selections and our needs. Some of the common types of visualizations that can be used are bar charts, line charts, maps etc. Each types of chart has their own uses and have to be used accordingly

HERE ARE SOME IMPORTANT VISUALIZATIONS:



The bar chart titled "COUNTRY BY NATIONALITY" from the "AIRLINES ANALYSIS" dataset provides the count of individuals by nationality. Here are the key insights from the graph:
United States Dominance: The United States has a significantly higher count of individuals compared to any other country, with the count exceeding 20,000. This indicates that the majority of the dataset's individuals are from the United States.
Other Major Nationalities:
Australia: The second highest, but much lower than the US, with a count between 5,000 and 10,000.
Canada and Brazil: Both have similar counts, slightly lower than Australia.
Papua New Guinea: Also has a notable count, similar to Canada and Brazil.
China and Indonesia: Have moderate counts, lower than the top five but still significant.
Mid-Range Nationalities:
Countries like the Russian Federation, Colombia, India, and France fall into the mid-range category, with counts decreasing gradually as we move down the list.
Lower-End Nationalities:
The United Kingdom, Argentina, Germany, Mexico, Japan, and other countries have smaller counts, each with less than 5,000 individuals.
This pattern continues with countries like South Africa, Iran, Philippines, Venezuela, Pakistan, Turkey, Malaysia, and others having even fewer individuals.
Least Represented Nationalities:
Countries such as Chile, Sweden, Ethiopia, French Polynesia, Thailand, Greece, Peru, Myanmar, and Algeria are at the lower end of the spectrum with the smallest counts in the dataset.
The graph indicates a heavy skew towards individuals from the United States, with a steep drop-off as we move to other nationalities. The distribution suggests that a few countries dominate the dataset while many others have relatively small representations.



The pie chart titled "GENDER DISTRIBUTION" from the "AIRLINES ANALYSIS" dataset shows the proportion of males and females. Here are the key insights from the graph:

Almost Equal Gender Distribution:

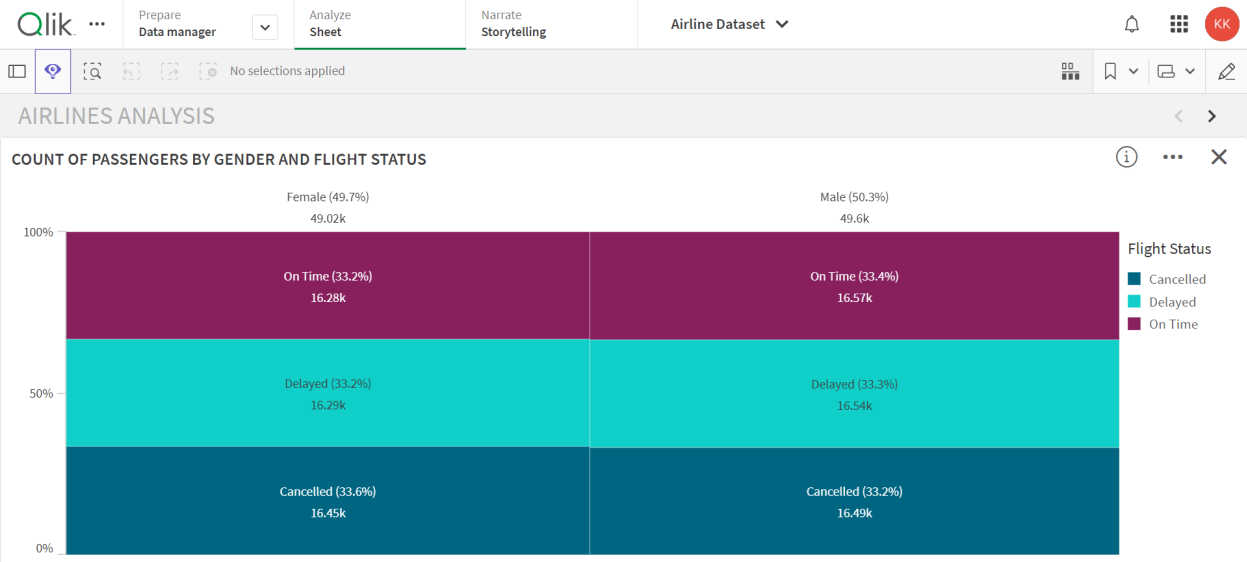
Males constitute 50.3% of the dataset.

Females make up 49.7% of the dataset.

Slight Male Majority:

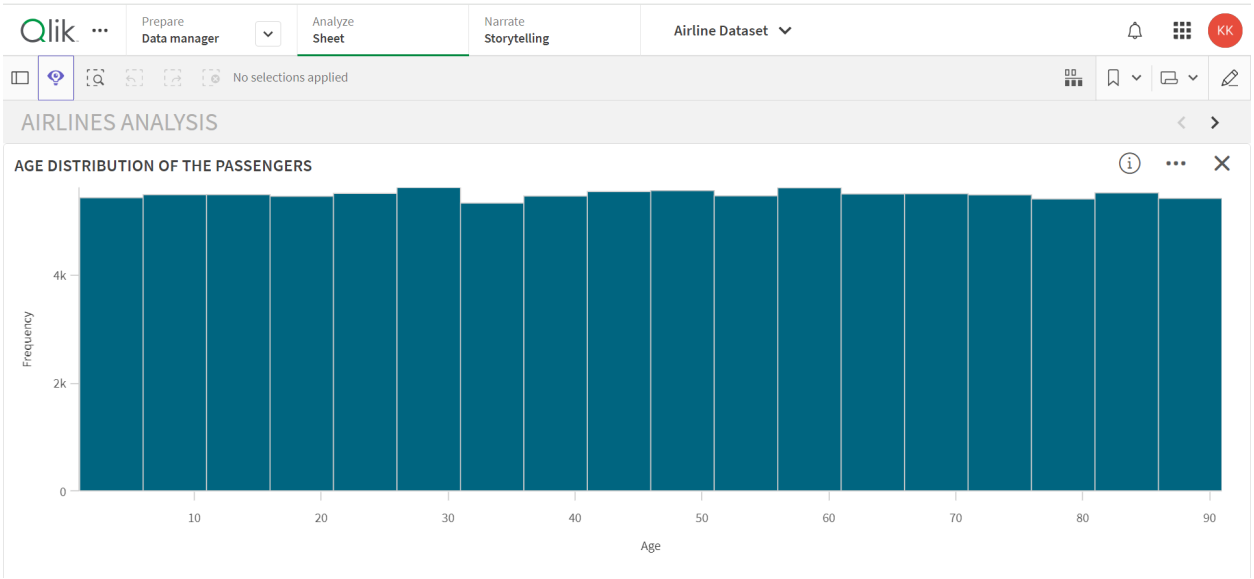
There is a very slight majority of males over females, with the difference being just 0.6%.

The distribution is almost evenly split between males and females, indicating a balanced representation of genders in the dataset.



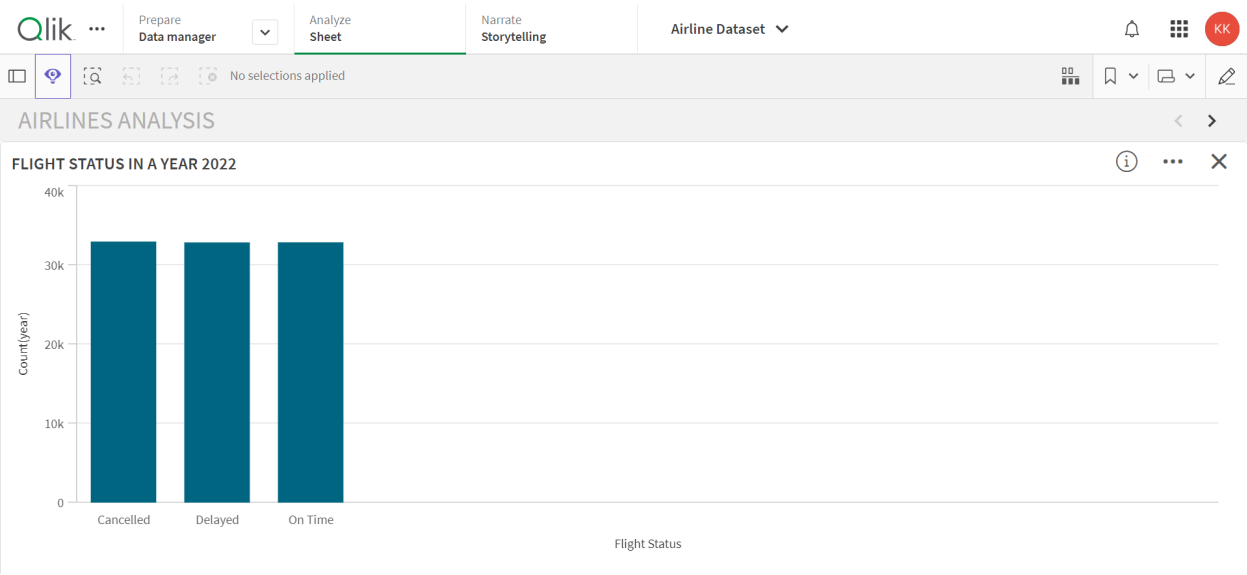
The bar chart titled "COUNT OF PASSENGERS BY GENDER AND FLIGHT STATUS" from the "AIRLINES ANALYSIS" dataset provides a breakdown of the flight status (On Time, Delayed, Cancelled) for male and female passengers. Here are the key insights from the graph:

Overall Gender Distribution:
The total number of male passengers is 49.6k (50.3%).
The total number of female passengers is 49.02k (49.7%).
Flight Status Distribution for Females:
On Time: 16.28k (33.2% of female passengers)
Delayed: 16.29k (33.2% of female passengers)
Cancelled: 16.45k (33.6% of female passengers)
Flight Status Distribution for Males:
On Time: 16.57k (33.4% of male passengers)
Delayed: 16.54k (33.3% of male passengers)
Cancelled: 16.49k (33.2% of male passengers)
Comparison of Flight Status Between Genders:
The percentages of on-time, delayed, and cancelled flights are almost identical for both genders.
Slightly more male flights are on time compared to female flights (33.4% vs. 33.2%).
The delayed and cancelled flight percentages are nearly equal between males and females.
Balanced Representation in Flight Status:
The flight status distribution (on time, delayed, cancelled) is very similar for both males and females, indicating no significant gender-based difference in flight status experiences.
Overall, the graph indicates a balanced representation of flight status across genders, with no significant disparity between males and females in terms of on-time, delayed, or cancelled flights.



The graph shows the age distribution of passengers. Here are some insights based on the graph:

Uniform Distribution:
The age distribution is quite uniform across different age groups. Each age group has a relatively similar frequency, indicating that passengers come from a wide range of age groups without any significant skew towards any particular age range.
Frequency Consistency:
The frequency of passengers in each age group is fairly consistent, mostly around the 4,000 mark. This suggests that the airline services are equally utilized by people of all age groups.
Slight Variations:
There are minor variations, but they are not very pronounced. For instance, the age groups around 10, 20, and 80 have slightly higher frequencies compared to others, but the differences are not substantial.
Inclusivity:
The data indicates that the airline caters to a diverse age demographic, ensuring services are accessible and appealing to both young and older passengers.
Overall, the graph suggests a well-balanced age distribution among passengers, with a consistent frequency across various age groups. This balance indicates no particular age group is disproportionately represented among the airline's passengers.



The graph title is "FLIGHT STATUS IN A YEAR 2022". The x-axis is labeled "Court(year)" but it likely refers to "Count(year)" which means the number of flights for a given year. The y-axis is labeled "Flight Status". There are three categories plotted on the y-axis: Cancelled, Delayed and On Time.

The data shows that in the year 2022, there were more cancelled flights than delayed flights, and the number of on-time flights was the highest. There were approximately 40,000 cancelled

flights, 30,000 delayed flights and 10,000 on-time flights.

Here are some additional insights that can be gleaned from the data:

- The cancellation rate for flights in 2022 was significantly higher than the delay rate. This could be due to a number of factors, such as bad weather, mechanical problems, or air traffic control issues.
- The on-time flight rate was the highest in 2022. This suggests that airlines were able to operate their flights efficiently for the most part.

It is important to note that the data in this graph is for the year 2022 only

Here are some additional questions that could be answered with more data:

- What were the most common reasons for flight cancellations and delays in 2022?
- How did flight performance vary by airline in 2022?
- How has flight performance changed over time?

Reasons for Cancellations and Delays:

- **Break down the "Cancelled" and "Delayed" categories:** The current graph lumps all cancellations and delays together. Further data could show subcategories like bad weather, mechanical issues, staffing shortages, air traffic control issues, and more. This breakdown would reveal the major contributors to disruptions.
- **Seasonal Variations:** Does the cancellation/delay rate fluctuate throughout the year? For example, bad weather might cause more cancellations in winter, while staffing shortages could be more prominent during peak travel seasons.

Performance by Airline:

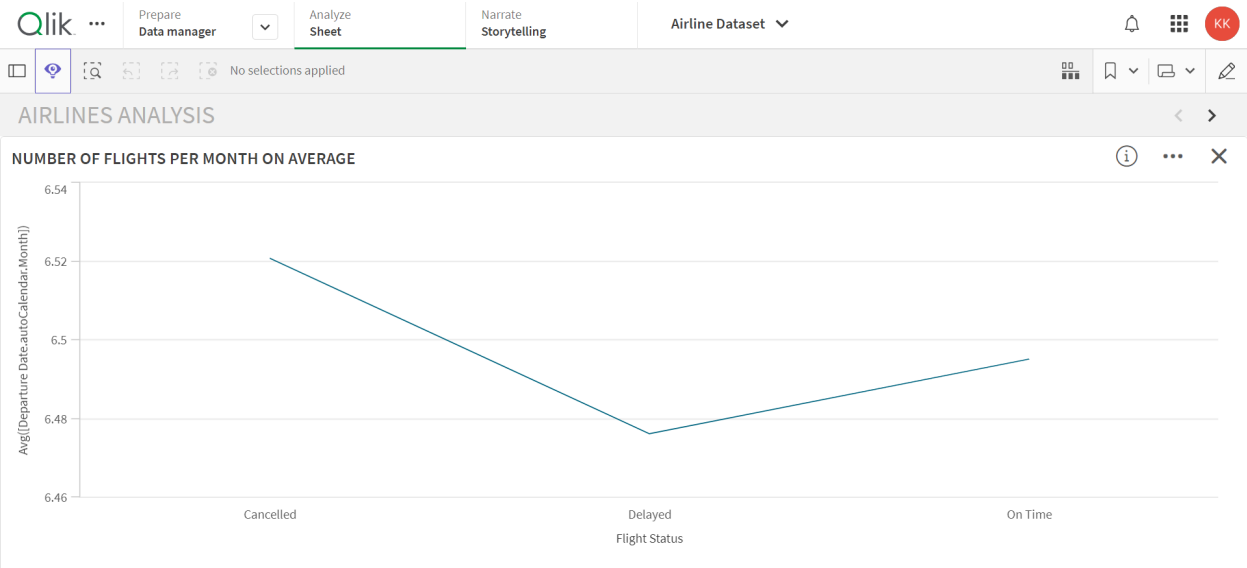
- **Compare Airlines:** How do different airlines stack up in terms of on-time performance, cancellation rates, and delay rates? This would allow travelers to make informed choices when booking flights.
- **Analyze Airline Trends:** Did any particular airlines experience significant changes in flight performance throughout 2022? Identifying these trends could point to specific challenges faced by those airlines.

Performance Over Time:

- **Historical Trends:** Adding data from previous years would show how the cancellation and delay rates have changed over time. Are cancellations becoming more or less frequent? Have airlines improved their on-time performance over the years?
- **External Factors:** Correlating flight performance data with external factors like major

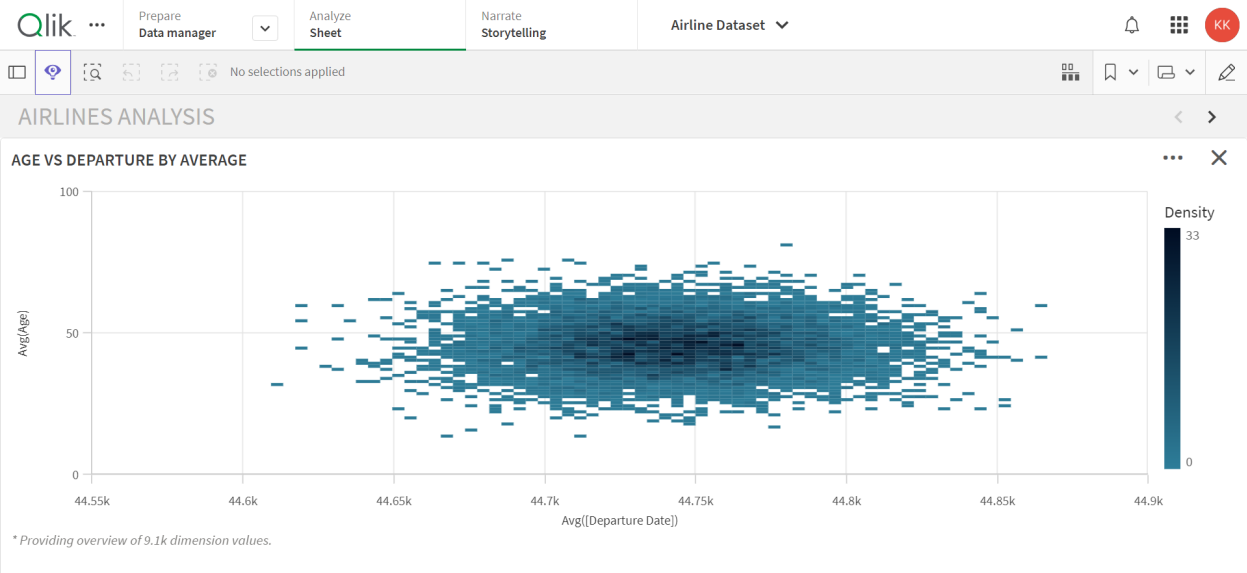
weather events, industry-wide staffing changes, or economic fluctuations could reveal broader trends impacting the entire airline industry.

By delving deeper with additional data, we can move beyond a basic understanding of cancellations and delays in 2022 and gain valuable insights into airline operations, travel planning strategies, and potential areas for improvement within the industry.



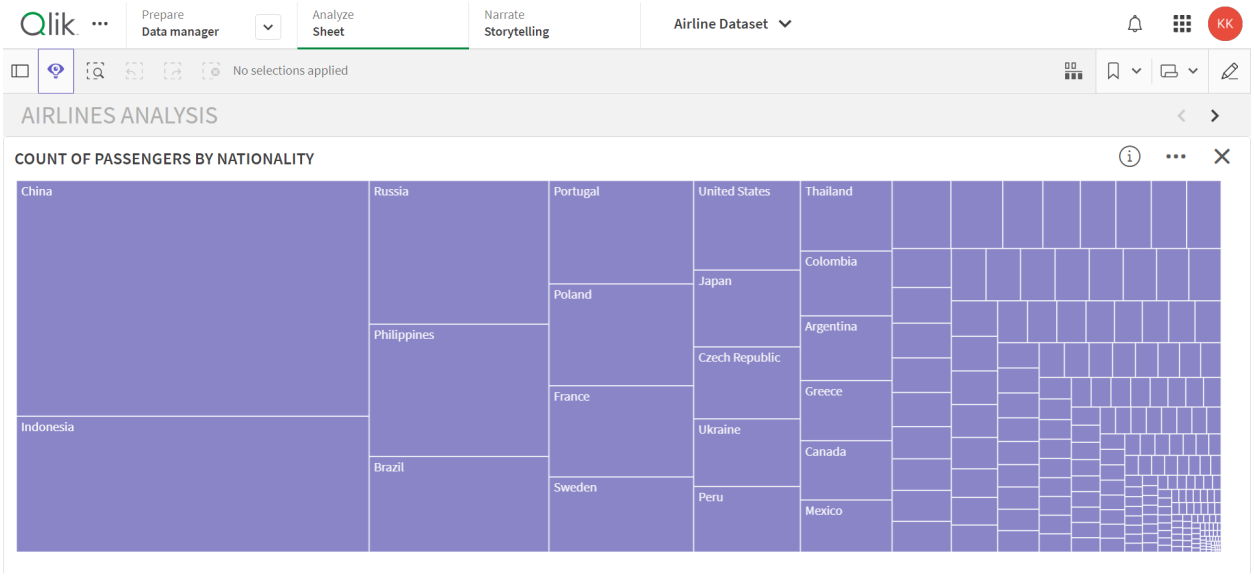
The graph provides a visualization of the average number of flights per month, categorized by flight status (Cancelled, Delayed, On Time). Here are some insights based on the graph:
Cancelled Flights:
The average number of cancelled flights per month is approximately 6.52. This represents the peak of the dataset.
Delayed Flights:
The average number of delayed flights per month drops to around 6.48. This is the lowest point on the graph, indicating that delayed flights occur less frequently on average compared to cancelled or on-time flights.
On-Time Flights:
The average number of on-time flights per month increases slightly to approximately 6.50. This is higher than delayed flights but lower than cancelled flights.
Trend Analysis:
There is a noticeable trend where cancelled flights are the most common, followed by on-time flights, and then delayed flights. This suggests that while there are disruptions causing cancellations, the majority of flights tend to be on time more often than delayed.
Monthly Flight Averages:

The graph shows a narrow range in the average number of flights per month (between 6.48 and 6.52), indicating relatively consistent flight frequencies regardless of their status.
The insights derived from the graph suggest that while cancellations are a significant issue, the majority of flights tend to be on time, with delays being the least frequent among the three categories.



The graph depicts a density plot of the average age of passengers versus the average departure date. Here are the insights based on the graph:
Concentration of Data Points:
There is a high concentration of data points around certain values, particularly between the ages of 30 and 70. This indicates that most passengers fall within this age range.
Uniformity Over Time:
The average departure dates are spread over a range (approximately 44.55k to 44.9k). Within this range, the density of passengers remains fairly consistent, suggesting no significant seasonal or temporal trends in the average age of passengers.
Age Distribution:
The graph shows that there are few passengers below the age of 20 and above the age of 80. The majority of the passengers are clustered in the middle age ranges.
Density Variations:
The density is highest around the average ages of 40 to 60, indicating that middle-aged passengers are the most common demographic.
Outliers:
There are some outliers with higher and lower ages, but these are relatively sparse. This suggests that while the airline serves passengers of all ages, those outside the 20 to 80 age range are less frequent.

Temporal Stability:
Over the period represented by the departure dates, there are no dramatic changes in the average age of passengers. This stability indicates a consistent customer base in terms of age over time.
Overall, the graph suggests that the airline's passenger demographic is relatively stable over time, with the majority of passengers being middle-aged. There are fewer very young or very old passengers, and the data does not indicate any significant temporal fluctuations in the age distribution.



the insights from the treemap graph titled “AIRLINES ANALYSIS BY NATIONALITY.”

China: The largest rectangle represents China, indicating that it has the highest count of passengers among the nationalities in this dataset. China’s air travel market is significant.

India: The next sizable rectangle is labeled “India,” suggesting a substantial number of passengers from India. India’s growing economy and increasing air connectivity contribute to

this trend.

Russia: Russia also stands out with a notable passenger count. It could be due to business travel, tourism, or other factors.

Portugal: The rectangle for Portugal is relatively large, indicating a significant number of passengers. Portugal's popularity as a tourist destination might play a role.

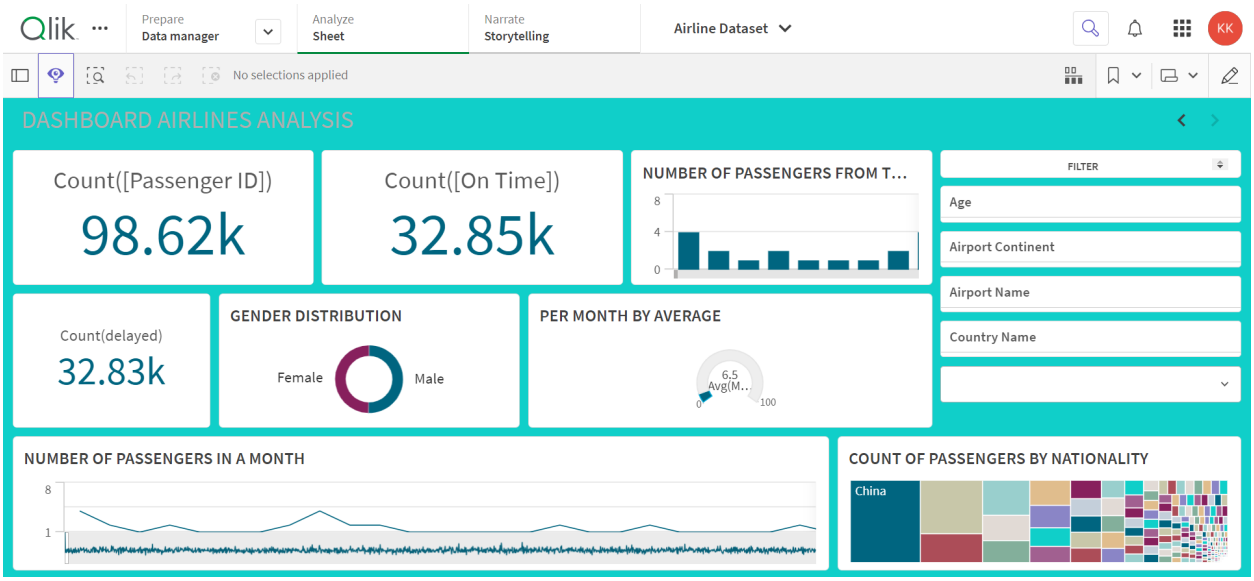
United States: Although not the largest, the U.S. rectangle is still substantial. The U.S. has a robust domestic and international air travel market.

Thailand, Colombia, Japan, and Argentina: These countries have moderate passenger counts. Tourism, business, and cultural exchanges likely contribute.

Czech Republic, Ukraine, and Mexico: Smaller rectangles represent these countries. While fewer passengers, they still contribute to global air travel.
Remember that this analysis is based on the available data, and other factors like airline routes, economic ties, and travel policies influence passenger numbers.

6.)DASHBOARDS

Qlik Sense Cloud is a powerful tool for creating interactive dashboards and data visualizations. Here are the steps and key features for creating and working with dashboards in Qlik Sense Cloud:



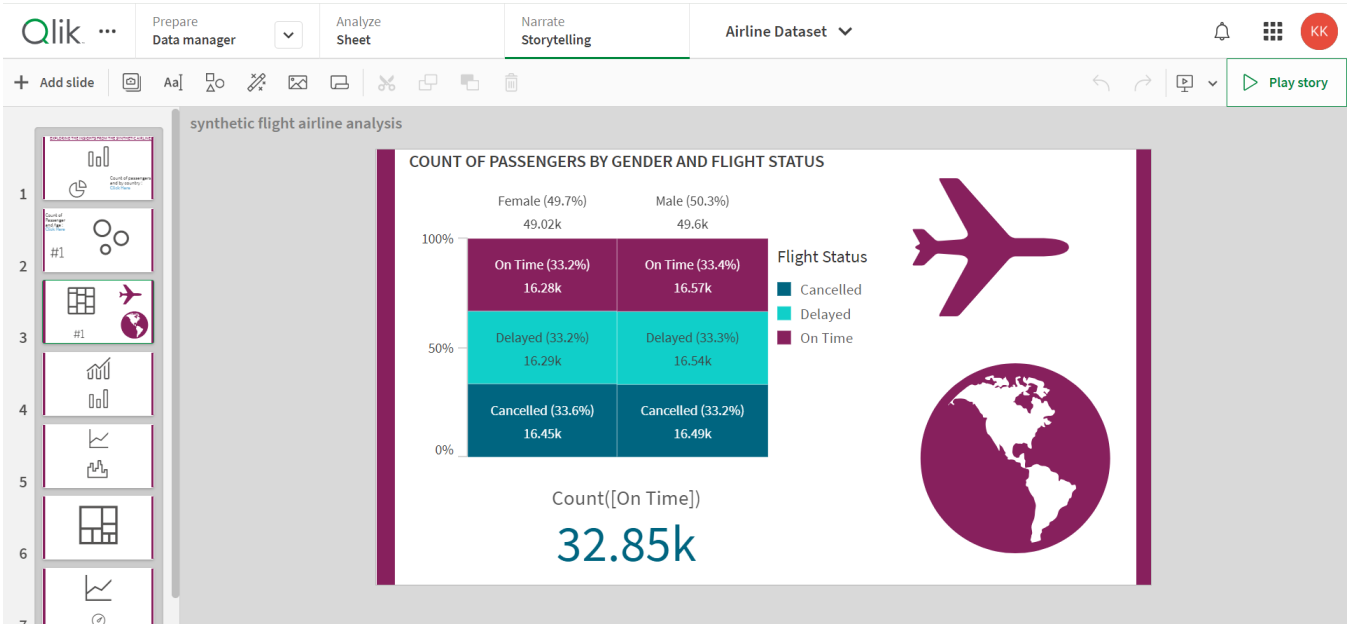
The dashboard presented provides a comprehensive analysis of airline passengers, focusing on various metrics and visualizations. Here are the insights derived from the dashboard:
Overview
<u>Total Passengers</u>:
The total number of passengers is 98.62k.

On-Time Flights:
The number of on-time flights is 32.85k.
Delayed Flights:
The number of delayed flights is 32.83k, which is very close to the number of on-time flights, indicating potential issues with flight punctuality.
Detailed Insights
Gender Distribution:
The gender distribution pie chart shows the proportion of male and female passengers. This helps understand the demographic split of the passengers.
Passengers per Month:
The gauge indicates the average number of passengers per month, which is 6.5. This metric helps in understanding the monthly passenger flow.
Number of Passengers by Nationality:
The treemap visualizes the number of passengers based on their nationality, highlighting China prominently. This indicates that a significant portion of passengers are from China.
Number of Passengers from Top Airports:
The bar chart shows the number of passengers from the top airports. This helps identify the busiest airports in terms of passenger traffic.
Number of Passengers in a Month:
The line chart at the bottom left shows the trend of the number of passengers over a month. This can help in identifying seasonal variations or peak travel times.
Filters
Age, Airport Continent, Airport Name, Country Name:
These filters on the right side allow users to slice and dice the data to focus on specific demographics, geographical regions, or airport-specific data.
Key Observations
Flight Delays:
The number of delayed flights being almost equal to the number of on-time flights suggests there could be operational challenges affecting punctuality.
Passenger Demographics:
Understanding the gender distribution helps in tailoring services and amenities to better cater to the needs of the predominant passenger gender.
Monthly Passenger Trends:
The average monthly passenger count provides insights into overall airline usage and can help in resource allocation and planning.
Nationality Distribution:
The prominence of Chinese passengers indicates a potentially significant market segment, which can inform targeted marketing and service improvements.
Busiest Airports:
Knowing the top airports by passenger count can help in logistical planning and

improving services at these key hubs.
By leveraging these insights, the airline can make data-driven decisions to improve operational efficiency, enhance passenger experience, and target marketing efforts more effectively.

7.) STORY TELLING

Storytelling is very much similar to the presentations we do with slides, images etc. As the name suggests, storytelling is used to tell 'stories' with the data, explaining it visually to people who might not be very well versed with the numbers and technical jargons being used at many places. Our visualizations and dashboards should be in such a way that even peopl who do not know much about things should be able to make out what is being said to them easily. Similarly, we use Qlik Sense's storytelling feature to make our visualizations into slides that explain what the charts are.



9.)PERFORMANCE TESTING:

AMOUNT OF DATA RENDERED:

"Amount of Data Loaded" refers to the quantity or volume of data that has been imported, retrieved, or loaded into a system, software application, database, or any other

data storage or processing environment. It's a measure of how much data has been successfully processed and made available for analysis, manipulation, or use within the system

we have uploaded 1 csv file in the dataset in the qlick sense used the dataset in our visualization. We have not removed or added any fields to our datasets. In that case, the data amounts to total of only around 50 KB, which is very lightweight and hence, it would be easier for Qlik Sense to work with this smaller size of data.

UTILIZATION OF FILTERS:

"Utilization of Filters" refers to the application or use of filters within a system, software application, or data processing pipeline to selectively extract, manipulate, or analyze data based on specified criteria or conditions. Filters are used to narrow down the scope of data, focusing only on the relevant information that meets certain predefined criteria

We have applied data filters wherever required to limit the amount of data shown on the charts. This is done if there are too many data and we only want to see only a specific number of them, or if we want to filter out any unwanted data