CRANE COMPANY Business Model Proposal

At CRANE COMPANY, we are committed to providing our customers with high-quality products and services. As part of our growth strategy, we are exploring opportunities to expand our operations into new markets. We believe that by expanding into new markets, we can increase our customer base, grow our revenue, and establish ourselves as a global leader in our industry.

Furthermore, we recognize that expanding our operations can lead to increased brand recognition and greater market share, as well as potential cost savings and other economies of scale. This written assignment will explore the problems, opportunities and benefits that we can gain through expansion, as well as the potential challenges and risks associated with this process.

Problem statements

Due to a potential networking lead, a new business partner has offered the opportunity to expand the company's operations to the United States. This presents a significant opportunity for the company to enter a new market and potentially increase their customer base and revenue. However, expanding to a new country can also pose several challenges, including regulatory compliance, cultural differences, and logistical issues. It is important for the company to carefully consider the risks and benefits of such an expansion before making any decisions.

Problem 1 : Risk to enter the United States

Problem statement

As a company that has only been operating in the United Kingdom, there is a lot of uncertainty surrounding this prospect. While expanding to the United States could bring many benefits, there are also many challenges and risks to consider

Understanding

To assess whether expanding to the United States is a viable option for the company, an inspection of the World Bank Database will be conducted to determine if the country possesses the necessary attributes for successful expansion.

Problem 2: No overseas operation experience

Problem statement

The company lacks experience in selling products internationally, which has resulted in a significant degree of uncertainty regarding the matter.The company lacks experience in selling products internationally, which has resulted in a significant degree of uncertainty regarding the matter.

Understanding

There are several logistic subjects that need to be carefully considered, including imports and exports, as well as lead times. These factors are crucial in determining the feasibility of operating in a new market and should be thoroughly evaluated before any decisions are made.

Problem 3 Where to enter

Problem statement

The United States is a vast country with 50 different states, each with its unique economy, demographics, and consumer behavior. As a result, there are many different business opportunities that exist depending on the state.

Understanding

To determine the best business opportunity in the US, it is essential to analyze data from different sources, including economic and demographic data, industry trends, and consumer behavior. This information can provide insights into the potential demand for a product or service in a particular state, the level of competition, and the overall market potential.

Results

After a comprehensive analysis of the data, the results have revealed significant insights for each of the problems under investigation. The analysis included data cleaning, manipulation, and visualization techniques to provide an in-depth understanding of the data. This report aims to present the findings of the analysis and draw meaningful conclusions for each problem.

After a comprehensive analysis of the data, the results have revealed significant insights for each of the problems under investigation. The analysis included data cleaning, manipulation, and visualization techniques to provide an in-depth understanding of the data. This chapter aims to present the findings of the analysis and draw meaningful conclusions for each problem.

Problem 1

After analyzing the data, it has been determined that the United States is a good alternative for the company to expand its operations.

The United States holds the second place in the countries that spend the most per person and is the third most populated country. This mix of high per capita spending and a large population confirms that the total customer segment is big enough for the company to consider expanding its operations.

While there are risks associated with entering a new market, the potential benefits of expanding to the United States outweigh these challenges.

By using the World Bank Database to evaluate the country's attributes for successful expansion, it is clear that the United States is a viable option for the company to consider.

Problem 2

Although the company lacks experience in overseas operations, the data suggests that with careful consideration and planning, it could successfully expand into the United States market.

After analyzing the data, it was found that the company has a unique opportunity to compete with existing competitors due to its location.

The competitors are based in Asia, while the UK is much closer to the United States, which would result in reduced lead times and shipping costs.

Additionally, the data revealed that the safest way to ship would be through the "Same Day" option, ensuring that the company's products reach the destination on time and in perfect condition.

Problem 3

To determine the best state for expansion, we conducted an analysis of economic and demographic data, industry trends, and consumer behavior in each state.

The data revealed that the state with the most significant potential for growth is California, due to its large population, high per capita income, and robust economy.

Additionally, we found that the safest and most efficient way to ship our products would be to use the "same day" shipping option, which will ensure that our products arrive quickly and in good condition. Overall, by focusing on California and utilizing the "same day" shipping option, we can position ourselves for success in the United States market.

Assumptions made

To ensure the completeness and accuracy of this report, certain assumptions were made during the analysis process. These assumptions were necessary due to the limited information available at the time of the report's development, and are outlined below for clarity and transparency.To ensure the completeness and accuracy of this report, certain assumptions were made during the analysis process. These assumptions were necessary due to the limited information available at the time of the report's development, and are outlined below for clarity and transparency.

Problem 1

1) It was asumed that the target country is United States, no other countries where taking as consideration 2) The year for the Worl Bank is asumed as the most relevant for the analysis 3) The column "HouseholdExpenditurePerCapita" is asumed to be the Expenditure per person in one year 4) It is asumed

that the bigger the population the best market size. It was not posible to funnel the total population to the target audience.

Problem 2

1) It is asumed that the colum "customer country" corresponds to the country where the items are sold and the "order country" is asumed to be the country where the seller is. 2) The dataset corresponds to an Asian e-commerce store "AcmeSports". This is the only data source used to analize the sport clothing industry. 3) It is asumed that the total sports clothes are under the filters : Department Name= "apparel" and Category Name = "Women's Clothing" and "Men's Clothing"

Problem 3

1) It is asumed that the colum "customer country" corresponds to the country where the items are sold and the "order country" is asumed to be the country where the seller is. 2) The dataset corresponds to an Asian e-commerce store "AcmeSports". This is the only data source used to analize the sport clothing industry. 3) It is asumed that the total sports clothes are under the filters : Department Name= "apparel" and Category Name = "Women's Clothing" and "Men's Clothing" 4) It is asumed that the more clients under the state, the better for the company to enter

Limitations

In most cases your answers to the business questions will have some limitations. They might for example not be generalizable, but only valid for a certain case. Describe any limitations your results have.

Problem 1

There are several external factors that should be considered when expanding sales and entering the United States market. Some of the key factors to consider include:

1) The current economic and political climate 2) Market demand and competition 3) Regulatory requirements 4) cultural preferences.

It is important to conduct thorough research and analysis of these factors to make informed decisions about the feasibility and potential success of entering the US market. Additionally, it may be useful to seek guidance from local experts or consultants who have knowledge and experience working in the US market.

Problem 2

it is important to conduct a cultural analysis that considers factors such as fashion trends, consumer preferences, and cultural values. Some key aspects to consider include:

1) Fashion trends: The US is a diverse country, and fashion trends can vary widely depending on the region and demographic. It is important to research current and upcoming fashion trends to ensure that your

clothes will appeal to the target market. 2) Consumer preferences: Understanding the preferences and needs of your target audience is critical. Factors to consider may include age, gender, lifestyle, and purchasing habits. 3) Cultural values: Cultural values can impact how consumers perceive and react to different products. For example, Americans tend to value individuality, self-expression, and comfort. It is important to understand these cultural values and ensure that your clothes align with them. 4) Sizing: Clothing sizes may vary between countries, so it is important to ensure that your clothes are sized appropriately for the US market.

By considering these factors and conducting thorough market research, you can gain a better understanding of whether your clothes are likely to sell in the US market.

Problem 3

In addition to knowing the most optimal states to enter the US market, there are several other aspects to consider. One important factor is the legal and regulatory requirements for doing business in the US, including state and federal laws and regulations. Taxes are also an important consideration, as state taxes can vary significantly from one state to another. It is also important to consider the competitive landscape and market saturation in the industry, as well as consumer behavior and preferences in the target market. Additionally, cultural and linguistic differences may need to be taken into account when developing marketing and sales strategies. Other factors that may be relevant include supply chain and logistics, infrastructure, and labor costs.

Data

In this section you need to describe the data used, its sources, data quality, data constraints and the results of your EDA. This is likely one of the longer sections as it needs to go into detail here. It is important that the reader of your report is able to follow your thinking. Any code cells need to be executable top-to-bottom.

For each dataset the following should be answered:

- Why was this dataset used?
- For which problems was it used?
- Data source including link/code to get the data. Timestamps if the data is a snapshot.
- EDA
    - Data quality
    - constraints on the data

Dataset 1

In this chapter, we will conduct an Exploratory Data Analysis (EDA) for the "DataCoSupplyChainDataset.csv" file. This analysis is being conducted in response to a requirement that seeks to describe the data used, its sources, data quality, data constraints, and the results of the EDA. This section is expected to be longer than others, as it requires a detailed explanation of our thought process, and any code cells should be executable from top-to-bottom.

In [114…

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns



#Opening all the files that are going to be used and manipulated
chart = pd.read_csv('DataCoSupplyChainDataset.csv',encoding = "ISO-8859-1")

print(f"""

{"-"*120}
\033[1mDataset Name\033[0m
    DataCoSupplyChainDataset.csv

\033[1mWhy was this dataset used?\033[0m
    After analizing all possible files, this one the one that gave more information about the in
    market that contains spesifically the sports apparel category.

\033[1mFor which problems was it used?\033[0m
    This file is being used to solve the first and second problem that are focused on the sports
    foreign countries.

\033[1mWhat is the data source\033[0m
    The information given under the column "Product Image", indicates that the information given
    comes from an asian, online store called "AcmeSports", wich for the purpose of the analysis 
    considered the main competitor.

""")

print(f"""
{"-"*120}

\033[1mEDA Analysis\033[0m

""")

# Check the dimensions of the dataset
print(chart.info())

print(f"""

{"-"*120}
\033[1mResult analysis\033[0m
- The dataset contains 180,519 rows and 56 columns.
- There are no missing values for any of the columns except for "Order Zipcode" and "Product Des
- The data types for the columns include: object, int64, float64.
- There are several columns that appear to contain categorical data, such as "Type", "Delivery S
   "Customer Segment", "Market", "Order Region", "Order State", "Order Status", and "Shipping Mo
- The dataset includes information about customer orders, such as the order date, order ID, cust
  product ID, sales, and profit.
- The dataset also includes information about customers, such as their location, email, and name
- There are several columns that contain redundant or unnecessary information, such as "Customer
  "Product Description", and "Product Status". These columns may be dropped from the dataset dur
""")

print(f"""
{"-"*120}

\033[1mNext Step\033[0m
```

```
In order to optimize our processing time and ensure that we are working only with relevant data,
we will be creating a new data frame to be used from now on. This new data frame will include on
the columns that are essential for our analysis, namely: "Type", "Delivery Status", "Category Na
"Customer Country", "Customer State", "Department Name", "Order Country", "Sales", "Order Profit
Order", and "Order Status".

We will also be removing any rows with null values, as these values do not provide any useful
information and can slow down our processing time. By creating this new data frame and removing
any unnecessary data, we can streamline our analysis and ensure that we are working with the mos
relevant and up-to-date information.

""")

#Filtering the usefull columns and droping the null values
chart_real = chart[['Type', 'Delivery Status', 'Category Name', 'Customer Country',\
                'Customer State','Department Name', 'Order Country', 'Sales',\
                'Order Profit Per Order', 'Order Status','Shipping Mode' ]].dropna()
#Filtering only values that come from the United States
chart_Filtered = chart_real.loc[chart['Customer Country'] == 'EE. UU.']

# Check the dimensions of the dataset
print(chart_Filtered.info())

print(f"""
{"-"*120}

\033[1mData quality\033[0m
The provided information is a Pandas DataFrame with 111,146 entries and 11 columns.
The dataset appears to be relatively large, with no null values in any of the columns.
The data types are appropriate for the data they represent, with numerical data types
for the 'Sales' and 'Order Profit Per Order' columns and object data types for the
remaining columns.

\033[1mConstraints on the data\033[0m
 - The data is not complete for all months of the year
 - The data used only has information from 2015 to 2017, But when filtering by sports
   clothes the information given is only for one year (2017)
 - The information provided represents sales only from a Online Store that operates in Alibaba
   there is no local information or other sorces to compare



\033[1mResult and Conclusion\033[0m
As a result of the filtering, now we have a new data frame with the following attributes:

 - It contains 180,519 rows and 10 columns, with a RangeIndex that goes from 0 to 180,518.
 - The data frame has 8 columns of object data type and 2 columns of float64 data type.
 - All the columns have non-null values, meaning that all the rows have values for each column.
 - The memory usage for this data frame is 13.8+ MB, which is relatively small and optimized
   for further processing. (from 77.1MB to 13.8MB)


In this chapter we have established the importance of the dataset and filtered it to a more
manageable size, optimizing it for faster processing time. After applying the necessary filters
and removing null values, we are left with a clean dataset that consists of 180,519 rows and
10 columns.

In the next chapter, we will proceed with the analysis of this dataset to extract meaningful
insights and make data-driven decisions.
""")
```

```
------------------------------------------------------------------------------------------------
------------------------
```

Dataset Name

DataCoSupplyChainDataset.csv

Why was this dataset used?

After analizing all possible files, this one the one that gave more information about the international market that contains spesifically the sports apparel category.

For which problems was it used?

This file is being used to solve the first and second problem that are focused on the sports apparel industry in foreign countries.

What is the data source

The information given under the column "Product Image", indicates that the information given comes from an asian, online store called "AcmeSports", wich for the purpose of the analysis will be considered the main competitor.

```
------------------------------------------------------------------------------------------------
------------------------
```

EDA Analysis

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 180519 entries, 0 to 180518
Data columns (total 56 columns):
 #   Column                         Non-Null Count   Dtype  
---  ------                         --------------   -----  
 0   Type                           180519 non-null  object 
 1   Days for shipping (real)       180519 non-null  int64  
 2   Days for shipment (scheduled)  180519 non-null  int64  
 3   Benefit per order              180519 non-null  float64
 4   Sales per customer             180519 non-null  float64
 5   Delivery Status                180519 non-null  object 
 6   Late_delivery_risk             180519 non-null  int64  
 7   Category Id                    180519 non-null  int64  
 8   Category Name                  180519 non-null  object 
 9   Customer City                  180519 non-null  object 
 10  Customer Country               180519 non-null  object 
 11  Customer Email                 180519 non-null  object 
 12  Customer Fname                 180519 non-null  object 
 13  Customer Id                    180519 non-null  int64  
 14  Customer Lname                 180511 non-null  object 
 15  Customer Password              180519 non-null  object 
 16  Customer Segment               180519 non-null  object 
 17  Customer State                 180519 non-null  object 
 18  Customer Street                180519 non-null  object 
 19  Customer Zipcode               180516 non-null  float64
 20  Department Id                  180519 non-null  int64  
 21  Department Name                180519 non-null  object 
 22  Latitude                       180519 non-null  float64
 23  Longitude                      180519 non-null  float64
 24  Market                         180519 non-null  object 
 25  Order City                     180519 non-null  object 
 26  Order Country                  180519 non-null  object 
 27  Order Customer Id              180519 non-null  int64
 28  order date (DateOrders)        180519 non-null  object
 29  Order Id                       180519 non-null  int64
 30  Order Item Cardprod Id         180519 non-null  int64
 31  Order Item Discount            180519 non-null  float64
 32  Order Item Discount Rate       180519 non-null  float64
 33  Order Item Id                  180519 non-null  int64
 34  Order Item Product Price       180519 non-null  float64
 35  Order Item Profit Ratio        180519 non-null  float64
 36  Order Item Quantity            180519 non-null  int64
 37  Sales                          180519 non-null  float64
 38  Order Item Total               180519 non-null  float64
 39  Order Profit Per Order         180519 non-null  float64
 40  Order Region                   180519 non-null  object
 41  Order State                    180519 non-null  object
 42  Order Status                   180519 non-null  object
 43  Order Zipcode                  24840 non-null   float64
 44  Product Card Id                180519 non-null  int64
 45  Product Category Id            180519 non-null  int64
 46  Product Description            0 non-null       float64
 47  Product Image                  180519 non-null  object
 48  Product Name                   180519 non-null  object
 49  Product Price                  180519 non-null  float64
 50  Product Status                 180519 non-null  int64
 51  shipping date (DateOrders)     180519 non-null  object
 52  Shipping Mode                  180519 non-null  object
 53  Year                           180519 non-null  int64
 54  Month                          180519 non-null  int64
 55  Day                            180519 non-null  int64
dtypes: float64(15), int64(17), object(24)
memory usage: 77.1+ MB
None
```

Result analysis

- The dataset contains 180,519 rows and 56 columns.
- There are no missing values for any of the columns except for "Order Zipcode" and "Product Description".
- The data types for the columns include: object, int64, float64.
- There are several columns that appear to contain categorical data, such as "Type", "Delivery Status",
   "Customer Segment", "Market", "Order Region", "Order State", "Order Status", and "Shipping Mode".
- The dataset includes information about customer orders, such as the order date, order ID, customer ID,
  product ID, sales, and profit.
- The dataset also includes information about customers, such as their location, email, and name.
- There are several columns that contain redundant or unnecessary information, such as "Customer Password",
  "Product Description", and "Product Status". These columns may be dropped from the dataset during preprocessing.

Next Step

In order to optimize our processing time and ensure that we are working only with relevant data, we will be creating a new data frame to be used from now on. This new data frame will include on ly the columns that are essential for our analysis, namely: "Type", "Delivery Status", "Category Name", "Customer Country", "Customer State", "Department Name", "Order Country", "Sales", "Order Profit Per Order", and "Order Status".

We will also be removing any rows with null values, as these values do not provide any useful information and can slow down our processing time. By creating this new data frame and removing any unnecessary data, we can streamline our analysis and ensure that we are working with the most relevant and up-to-date information.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 111146 entries, 2 to 180516
Data columns (total 11 columns):
 #   Column                  Non-Null Count   Dtype  
---  ------                  --------------   -----  
 0   Type                    111146 non-null  object 
 1   Delivery Status         111146 non-null  object 
 2   Category Name           111146 non-null  object 
 3   Customer Country        111146 non-null  object 
 4   Customer State          111146 non-null  object 
 5   Department Name         111146 non-null  object 
 6   Order Country           111146 non-null  object 
 7   Sales                   111146 non-null  float64
 8   Order Profit Per Order  111146 non-null  float64
 9   Order Status            111146 non-null  object 
 10  Shipping Mode           111146 non-null  object 
dtypes: float64(2), object(9)
memory usage: 10.2+ MB
None
```

```
------------------------------------------------------------------------------------------------------------------------
```

Data quality
The provided information is a Pandas DataFrame with 111,146 entries and 11 columns.
The dataset appears to be relatively large, with no null values in any of the columns.
The data types are appropriate for the data they represent, with numerical data types
for the 'Sales' and 'Order Profit Per Order' columns and object data types for the
remaining columns.

Constraints on the data

- The data is not complete for all months of the year
- The data used only has information from 2015 to 2017, But when filtering by sports clothes the information given is only for one year (2017)
- The information provided represents sales only from a Online Store that operates in Alibaba there is no local information or other sorces to compare

Result and Conclusion
As a result of the filtering, now we have a new data frame with the following attributes:

- It contains 180,519 rows and 10 columns, with a RangeIndex that goes from 0 to 180,518.
- The data frame has 8 columns of object data type and 2 columns of float64 data type.
- All the columns have non-null values, meaning that all the rows have values for each column.
- The memory usage for this data frame is 13.8+ MB, which is relatively small and optimized for further processing. (from 77.1MB to 13.8MB)

In this chapter we have established the importance of the dataset and filtered it to a more manageable size, optimizing it for faster processing time. After applying the necessary filters and removing null values, we are left with a clean dataset that consists of 180,519 rows and 10 columns.

In the next chapter, we will proceed with the analysis of this dataset to extract meaningful insights and make data-driven decisions.

Dataset 2

In [127…

```
#Opening all the files that are going to be used and manipulated
population = pd.read_csv('population_by_country_2020.csv')

print(f"""

{"-"*120}
\033[1mDataset Name\033[0m
    population_by_country_2020.csv

\033[1mWhy was this dataset used?\033[0m
    This file is part of the World Bank dataset, and it will play a crusial part for this projec
    file that will show whether the or not the United States has a population worth selling our

\033[1mFor which problems was it used?\033[0m
    This file is being used to solve the first problem regarding demographic characteristics.

\033[1mWhat is the data source\033[0m
    World Bank

""")

print(f"""
{"-"*120}

\033[1mEDA Analysis\033[0m

""")

print(population.head())

print(f"""

{"-"*120}
\033[1mResult analysis\033[0m
as we can see, the file contains categorical information regarding the demographic distribution
per country. For this excersice it will be used to cross the total population and compare it
with the other countries.

{"-"*120}

""")

print(population.info())

print(f"""
{"-"*120}
\033[1mDescriptive Analysis\033[0m
Now that we have the total dataset info, we will priceed to do a descriptive analisys for the co
""")
```

```
print(population['Population (2020)'].describe())

print(f"""
{"-"*120}
\033[1mResult analysis\033[0m
The result shows descriptive statistics for the "Population (2020)" column, which has a count of
The mean population is 33,227,440, with a standard deviation of 135,303,400, indicating a large
in population size among the countries.

The minimum population is 801, and the maximum is 1,440,298,000
showing a significant disparity between the least and most populous countries in the dataset. Th
percentile is 399,490, and the 75th percentile is 20,671,700, indicating that 50% of the countri
the dataset have a population between these two values.

These statistics provide a general overview of
the distribution of population sizes in the dataset, which can be used to inform further analysi
decision-making.
""")
print(f"""

{"-"*120}

\033[1mData quality\033[0m
The dataset used is limited and has only information for each country in idividual for an specif
In other terms is a snapshot and it does not necessarily indicates a true value.

\033[1mConstraints on the data\033[0m
 - The data is only for each country, it doesn't go further into details
 - The data does not indicate in what year it was extracted


\033[1mResult and Conclusion\033[0m
The data types of the columns in the dataframe are as follows: float64(1), int64(4), and object(
The columns with float64 and int64 data types likely represent numerical data such as population
area, and density, while the columns with object data types may contain text or mixed data types

The provided data frame has 235 entries and only 11 columns, making it relatively small compared
to other data sets. Due to its size and the absence of apparent inconsistencies in the data, it
is not necessary to filter or manipulate the data frame prior to analysis.

The data types of the columns seem appropriate for the information they contain, including integ
float, and object types. Therefore, the data frame can be directly used for analysis, and insigh
can be extracted to make informed decisions. However, it is still important to thoroughly examin
the data and ensure its accuracy and completeness before using it for any critical decision-maki


In the next chapter, we will proceed with the analysis of this dataset to extract meaningful
insights and make data-driven decisions.


{"-"*120}





""")
```


Dataset Name

population_by_country_2020.csv

Why was this dataset used?

This file is part of the World Bank dataset, and it will play a crusial part for this project because its the
file that will show whether the or not the United States has a population worth selling our products.

For which problems was it used?

This file is being used to solve the first problem regarding demographic characteristics.

What is the data source

World Bank

EDA Analysis

```
  Country (or dependency)  Population (2020) Yearly Change  Net Change  \
0                   China         1440297825        0.39 %     5540090   
1                   India         1382345085        0.99 %    13586631   
2           United States          331341050        0.59 %     1937734   
3               Indonesia          274021604        1.07 %     2898047   
4                Pakistan          221612785        2.00 %     4327022   

   Density (P/Km²)  Land Area (Km²)  Migrants (net) Fert. Rate Med. Age  \
0              153          9388211       -348399.0        1.7       38   
1              464          2973190       -532687.0        2.2       28   
2               36          9147420        954806.0        1.8       38   
3              151          1811570        -98955.0        2.3       30   
4              287           770880       -233379.0        3.6       23   

  Urban Pop % World Share  
0       61 %     18.47 %  
1       35 %     17.70 %  
2       83 %      4.25 %  
3       56 %      3.51 %  
4       35 %      2.83 %  
```


Result analysis

as we can see, the file contains categorical information regarding the demographic distribution per country. For this excersice it will be used to cross the total population and compare it with the other countries.


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 235 entries, 0 to 234
Data columns (total 11 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Country (or dependency)  235 non-null    object
 1   Population (2020)        235 non-null    int64
 2   Yearly Change            235 non-null    object
 3   Net Change               235 non-null    int64
 4   Density (P/Km²)          235 non-null    int64
 5   Land Area (Km²)          235 non-null    int64
 6   Migrants (net)           201 non-null    float64
 7   Fert. Rate               235 non-null    object
 8   Med. Age                 235 non-null    object
 9   Urban Pop %              235 non-null    object
 10  World Share              235 non-null    object
dtypes: float64(1), int64(4), object(6)
memory usage: 20.3+ KB
None
```

Descriptive Analysis

Now that we have the total dataset info, we will priceed to do a descriptive analisys for the column to be used

```
count    2.350000e+02
mean     3.322744e+07
std      1.353034e+08
min      8.010000e+02
25%      3.994905e+05
50%      5.460109e+06
75%      2.067170e+07
max      1.440298e+09
Name: Population (2020), dtype: float64
```

Result analysis

The result shows descriptive statistics for the "Population (2020)" column, which has a count of 235.
The mean population is 33,227,440, with a standard deviation of 135,303,400, indicating a large range
in population size among the countries.

The minimum population is 801, and the maximum is 1,440,298,000
showing a significant disparity between the least and most populous countries in the dataset. The 25th
percentile is 399,490, and the 75th percentile is 20,671,700, indicating that 50% of the countries in
the dataset have a population between these two values.

These statistics provide a general overview of
the distribution of population sizes in the dataset, which can be used to inform further analysis and
decision-making.

Data quality

The dataset used is limited and has only information for each country in idividual for an specific year.

```
In other terms is a snapshot and it does not necessarily indicates a true value.

Constraints on the data
 - The data is only for each country, it doesn't go further into details
 - The data does not indicate in what year it was extracted


Result and Conclusion
The data types of the columns in the dataframe are as follows: float64(1), int64(4), and object
(6).
The columns with float64 and int64 data types likely represent numerical data such as populatio
n,
area, and density, while the columns with object data types may contain text or mixed data type
s.

The provided data frame has 235 entries and only 11 columns, making it relatively small compared
to other data sets. Due to its size and the absence of apparent inconsistencies in the data, it
is not necessary to filter or manipulate the data frame prior to analysis.

The data types of the columns seem appropriate for the information they contain, including integ
er,
float, and object types. Therefore, the data frame can be directly used for analysis, and insigh
ts
can be extracted to make informed decisions. However, it is still important to thoroughly examin
e
the data and ensure its accuracy and completeness before using it for any critical decision-maki
ng.


In the next chapter, we will proceed with the analysis of this dataset to extract meaningful
insights and make data-driven decisions.


------------------------------------------------------------------------------------------------
------------------------
```

Dataset 3

In [129…

```
#Opening all the files that are going to be used and manipulated
hhe = pd.read_csv('householdexpenditure.csv')

print(f"""

{"-"*120}
\033[1mDataset Name\033[0m
    householdexpenditure.csv

\033[1mWhy was this dataset used?\033[0m
    This file is part of the World Bank dataset, and it will play a crusial part for this projec
    file that will show whether the or not the United States has a population worth selling our

\033[1mFor which problems was it used?\033[0m
    This file is being used to solve the first problem regarding demographic characteristics.

\033[1mWhat is the data source\033[0m
```

```
    World Bank

""")

print(f"""
{"-"*120}

\033[1mEDA Analysis\033[0m

""")

print(hhe.head())

print(f"""

{"-"*120}
\033[1mResult analysis\033[0m
As we can see, the file contains only two columns, one that stablishes the
country and the other that indicates the totl consuption per person.

this will play in important role for the analysis to determine if the United
States is a good idea.

{"-"*120}

""")
print(population.info())

print(f"""
{"-"*120}

\033[1mDescriptive Analysis\033[0m
Now that we have the total dataset info, we will priceed to do a descriptive analisys for the co
""")


print(population['Population (2020)'].describe())

print(f"""
{"-"*120}
\033[1mResult analysis\033[0m
The "Population (2020)" column contains 235 entries with a mean population of approximately
33.2 million and a standard deviation of 135.3 million.
The minimum population in the dataset is 801, while the maximum population is approximately
1.44 billion.

The first quartile (25th percentile) of the dataset is 399,490, while the median (50th percentil
) is 5.46 million, and the third quartile (75th percentile) is 20.67 million.

Overall, the population data appears to be widely varied, with a large range of values and a
relatively high standard deviation.
""")
print(hhe.info())

print(f"""

{"-"*120}
\033[1mResult and Conclusion\033[0m

\033[1mData quality\033[0m
The dataset used is limited and has only information for each country in idividual for an specif
In other terms is a snapshot and it does not necessarily indicates a true value.
```

```
\033[1mConstraints on the data\033[0m
 - The data is only for each country, it doesn't go further into details
 - The data does not indicate in what year it was extracted


The data types of the columns in the dataframe are as follows: float64(1), int64(4), and object(
The columns with float64 and int64 data types likely represent numerical data such as population
area, and density, while the columns with object data types may contain text or mixed data types

The provided data frame has 235 entries and only 11 columns, making it relatively small compared
to other data sets. Due to its size and the absence of apparent inconsistencies in the data, it
is not necessary to filter or manipulate the data frame prior to analysis.

The data types of the columns seem appropriate for the information they contain, including integ
float, and object types. Therefore, the data frame can be directly used for analysis, and insigh
can be extracted to make informed decisions. However, it is still important to thoroughly examin
the data and ensure its accuracy and completeness before using it for any critical decision-maki


In the next chapter, we will proceed with the analysis of this dataset to extract meaningful
insights and make data-driven decisions.


{"-"*120}





""")
```

```
-----------------------------------------------------------------------------------------------------------------------
```

Dataset Name

householdexpenditure.csv

Why was this dataset used?

This file is part of the World Bank dataset, and it will play a crusial part for this project because its the
file that will show whether the or not the United States has a population worth selling our products.

For which problems was it used?

This file is being used to solve the first problem regarding demographic characteristics.

What is the data source

World Bank

```
-----------------------------------------------------------------------------------------------------------------------
```

EDA Analysis

```
        Country  HouseholdExpenditurePerCapita
0     Hong Kong                          38285
1           USA                          37903
2   Switzerland                          28320
3    Luxembourg                          28261
4        Norway                          25481
```

```
-----------------------------------------------------------------------------------------------------------------------
```

Result analysis

As we can see, the file contains only two columns, one that stablishes the
country and the other that indicates the totl consuption per person.

this will play in important role for the analysis to determine if the United
States is a good idea.

```
-----------------------------------------------------------------------------------------------------------------------
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 235 entries, 0 to 234
Data columns (total 11 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   Country (or dependency)  235 non-null    object 
 1   Population (2020)        235 non-null    int64  
 2   Yearly Change            235 non-null    object 
 3   Net Change               235 non-null    int64  
 4   Density (P/Km²)          235 non-null    int64  
 5   Land Area (Km²)          235 non-null    int64  
 6   Migrants (net)           201 non-null    float64
 7   Fert. Rate               235 non-null    object 
 8   Med. Age                 235 non-null    object 
 9   Urban Pop %              235 non-null    object 
 10  World Share               235 non-null    object
dtypes: float64(1), int64(4), object(6)
memory usage: 20.3+ KB
None
```

```
------------------------------------------------------------------------------------------------
------------------------
```

Descriptive Analysis

Now that we have the total dataset info, we will priceed to do a descriptive analisys for the column to be used

```
count    2.350000e+02
mean     3.322744e+07
std      1.353034e+08
min      8.010000e+02
25%      3.994905e+05
50%      5.460109e+06
75%      2.067170e+07
max      1.440298e+09
Name: Population (2020), dtype: float64
```

```
------------------------------------------------------------------------------------------------
------------------------
```

Result analysis

The "Population (2020)" column contains 235 entries with a mean population of approximately 33.2 million and a standard deviation of 135.3 million.
The minimum population in the dataset is 801, while the maximum population is approximately 1.44 billion.

The first quartile (25th percentile) of the dataset is 399,490, while the median (50th percentile) is 5.46 million, and the third quartile (75th percentile) is 20.67 million.

Overall, the population data appears to be widely varied, with a large range of values and a relatively high standard deviation.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 176 entries, 0 to 175
Data columns (total 2 columns):
 #   Column                       Non-Null Count  Dtype 
---  ------                       --------------  ----- 
 0   Country                      176 non-null    object
 1   HouseholdExpenditurePerCapita  176 non-null    int64 
dtypes: int64(1), object(1)
memory usage: 2.9+ KB
None
```

```
------------------------------------------------------------------------------------------------
------------------------
```

Result and Conclusion

Data quality

The dataset used is limited and has only information for each country in idividual for an specific year.
In other terms is a snapshot and it does not necessarily indicates a true value.

Constraints on the data

 - The data is only for each country, it doesn't go further into details
 - The data does not indicate in what year it was extracted

```
The data types of the columns in the dataframe are as follows: float64(1), int64(4), and object
(6).
The columns with float64 and int64 data types likely represent numerical data such as populatio
n,
area, and density, while the columns with object data types may contain text or mixed data type
s.

The provided data frame has 235 entries and only 11 columns, making it relatively small compared
to other data sets. Due to its size and the absence of apparent inconsistencies in the data, it
is not necessary to filter or manipulate the data frame prior to analysis.

The data types of the columns seem appropriate for the information they contain, including integ
er,
float, and object types. Therefore, the data frame can be directly used for analysis, and insigh
ts
can be extracted to make informed decisions. However, it is still important to thoroughly examin
e
the data and ensure its accuracy and completeness before using it for any critical decision-maki
ng.


In the next chapter, we will proceed with the analysis of this dataset to extract meaningful
insights and make data-driven decisions.


------------------------------------------------------------------------------------------------
------------------------
```

Problem Solving

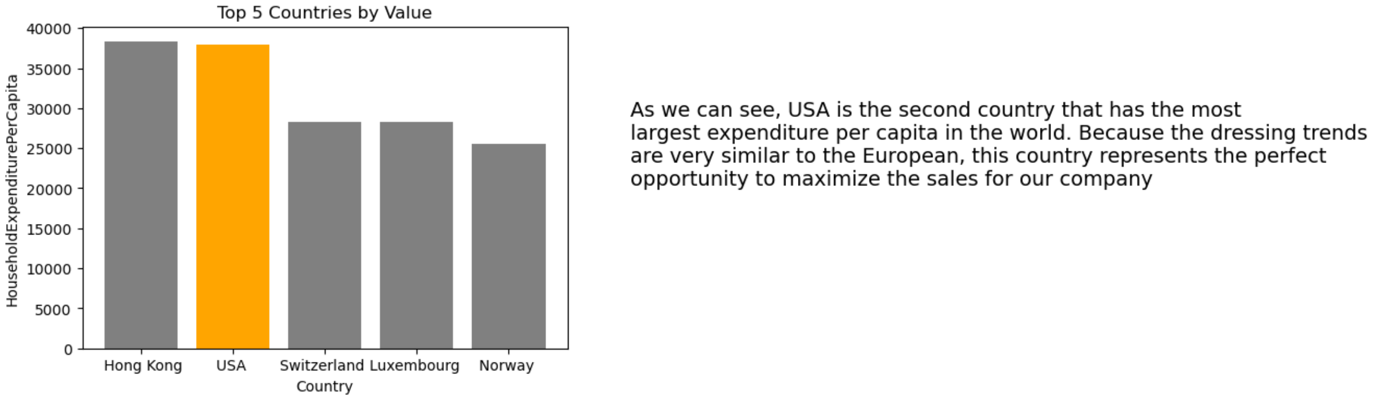
This section needs to guide throught the problem solving process and make it clear how the results have been derived from the data. It should also contain executable code for everything that is code based. Code cells need to be executable top-to-bottom and be well commented.

Problem 1

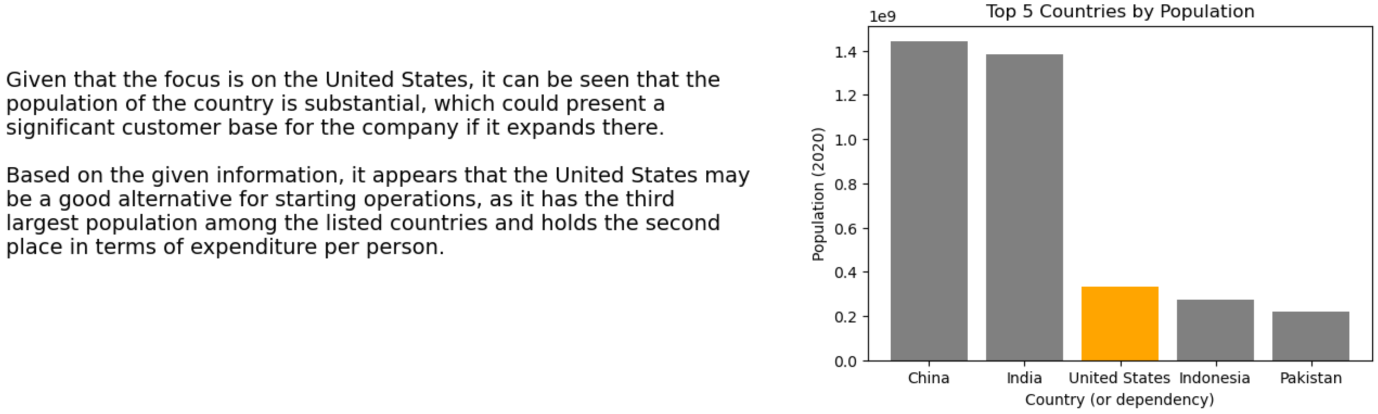
In [133…

```python
#creating a bar chart the biggest spending countries per person
hhe_sorted = hhe.sort_values(by='HouseholdExpenditurePerCapita', ascending=False)
hhe_top5 = hhe_sorted.head(5)

colors = ['grey', 'orange', 'grey', 'grey', 'grey']

plt.figure(figsize=(6, 4))
plt.bar(hhe_top5['Country'], hhe_top5['HouseholdExpenditurePerCapita'], color = colors)
plt.xlabel('Country')
plt.ylabel('HouseholdExpenditurePerCapita')
plt.title('Top 5 Countries by Value')
plt.text(1, 0.5, """As we can see, USA is the second country that has the most
largest expenditure per capita in the world. Because the dressing trends
are very similar to the European, this country represents the perfect
```

```
opportunity to maximize the sales for our company""",fontsize=14, transform=plt.gcf().transFigur



#Creating a bar chart for the top 5 biggest countries
population_sorted = population.sort_values(by= 'Population (2020)', ascending = False)
population_top5 = population_sorted.head(5)

colors = ['grey', 'grey', 'orange', 'grey', 'grey']

plt.figure(figsize=(6, 4))
plt.bar(population_top5['Country (or dependency)'], population_top5['Population (2020)'], color
plt.xlabel('Country (or dependency)')
plt.ylabel('Population (2020)')
plt.title('Top 5 Countries by Population')

plt.text(-1.2, 0.3, """
Given that the focus is on the United States, it can be seen that the
population of the country is substantial, which could present a
significant customer base for the company if it expands there.

Based on the given information, it appears that the United States may
be a good alternative for starting operations, as it has the third
largest population among the listed countries and holds the second
place in terms of expenditure per person.
""",fontsize=14, transform=plt.gcf().transFigure)



plt.show()
```



As we can see, USA is the second country that has the most largest expenditure per capita in the world. Because the dressing trends are very similar to the European, this country represents the perfect opportunity to maximize the sales for our company



Given that the focus is on the United States, it can be seen that the population of the country is substantial, which could present a significant customer base for the company if it expands there.

Based on the given information, it appears that the United States may be a good alternative for starting operations, as it has the third largest population among the listed countries and holds the second place in terms of expenditure per person.

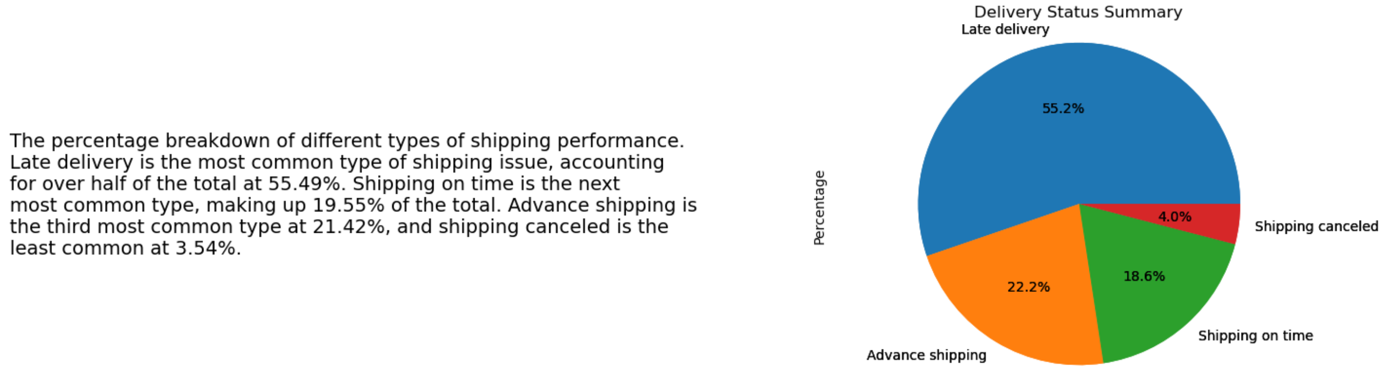
Problem 2

In [131…

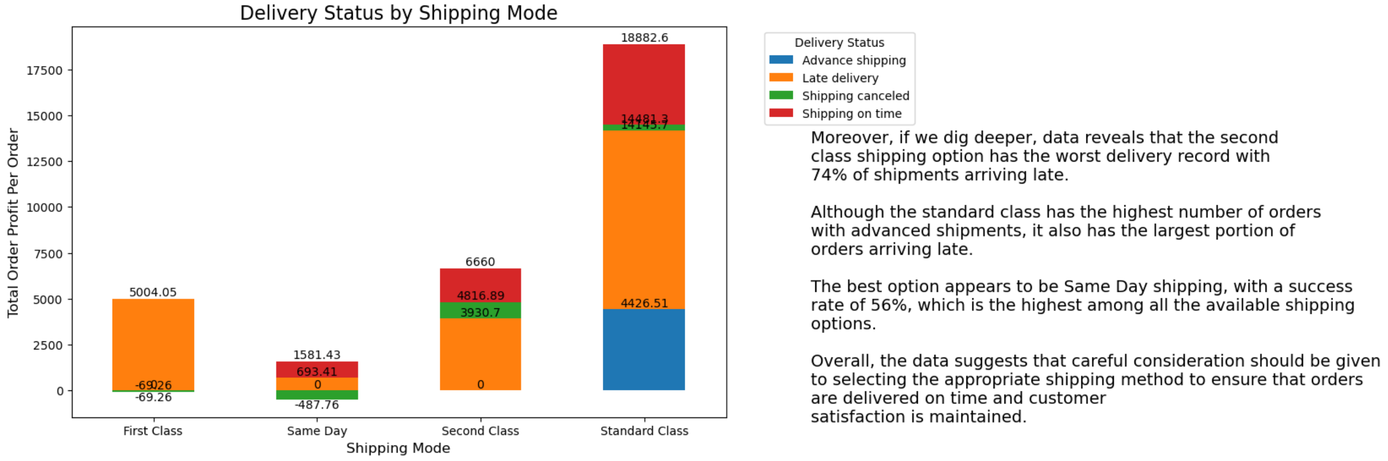
```
#--------------------------- # graph #1: Plot Pie-------------------------------------------
```

```
def summarize_delivery_status(df):
    delivery_status_counts = df['Delivery Status'].value_counts()
    delivery_status_percents = delivery_status_counts / delivery_status_counts.sum() * 100
    summary_delivery_status = pd.concat([delivery_status_counts, delivery_status_percents],
                                        axis=1, keys=['Count', 'Percentage'])
    return summary_delivery_status

#creating a new DataFrame to be used for the pie chart plot.
delivery_status = chart_Filtered.loc[(chart_Filtered['Department Name'] == 'Apparel') & \
                            (chart_Filtered['Category Name'].isin(["Women's Clothing",
                                                          "Men's Clothing",
                                                          "Children's Clothing"]))]

summary = summarize_delivery_status(delivery_status)

# Create a pie chart
summary['Percentage'].plot.pie(autopct='%1.1f%%')
summary['Percentage'].plot.pie(autopct='%1.1f%%')
plt.axis('equal')
plt.title('Delivery Status Summary')
plt.text(-1.2, 0.3, """

The percentage breakdown of different types of shipping performance.
Late delivery is the most common type of shipping issue, accounting
for over half of the total at 55.49%. Shipping on time is the next
most common type, making up 19.55% of the total. Advance shipping is
the third most common type at 21.42%, and shipping canceled is the
least common at 3.54%.

""",fontsize=14, transform=plt.gcf().transFigure)
plt.show()

#-------------------------------#Graph #2 Bar Chart, delivery status----------------------------

def summarize_delivery_status(df):
    shipping_mode = 'Shipping Mode'
    delivery_status = 'Delivery Status'

    summary = pd.pivot_table(df,
                             values='Order Profit Per Order',
                             index=[shipping_mode],
                             columns=[delivery_status],
                             aggfunc='sum',
                             fill_value=0)

    summary[delivery_status + ' %'] = summary.sum(axis=1) / summary.sum().sum() * 100
    return summary

pivot = delivery_status.pivot_table(index='Shipping Mode', columns='Delivery Status'\
                                    , values='Order Profit Per Order', aggfunc='sum')

# Create stacked bar plot
ax = pivot.plot(kind='bar', stacked=True, figsize=(10, 6))

# Add labels to each bar
for i in ax.containers:
    ax.bar_label(i, label_type='edge')

# Set plot properties

plt.title('Delivery Status by Shipping Mode', fontsize=16)
plt.xlabel('Shipping Mode', fontsize=12)
plt.xticks(rotation=0)
plt.ylabel('Total Order Profit Per Order', fontsize=12)
plt.legend(title='Delivery Status', bbox_to_anchor=(1.05, 1), loc='upper left')

txt = """
Moreover, if we dig deeper, data reveals that the second
class shipping option has the worst delivery record with
74% of shipments arriving late.

Although the standard class has the highest number of orders
with advanced shipments, it also has the largest portion of
orders arriving late.

The best option appears to be Same Day shipping, with a success
rate of 56%, which is the highest among all the available shipping
options.

Overall, the data suggests that careful consideration should be given
to selecting the appropriate shipping method to ensure that orders
are delivered on time and customer
satisfaction is maintained."""

plt.text(1, 0.1, txt, fontsize=14, transform=plt.gcf().transFigure)

plt.show()


#------------------------------ # Graph #3: Bar chart for suppliers-----------------------------

def summarize_order_country(df, department_name, customer_country, category_names):
    delivery_status = df.loc[(df['Department Name'] == department_name) & \
                             (df['Customer Country'] == customer_country) & \
                             (df['Category Name'].isin(category_names))]
    order_country_counts = delivery_status['Order Country'].value_counts()
    order_country_percents = order_country_counts / order_country_counts.sum() * 100
    summary1 = pd.concat([order_country_counts, order_country_percents], axis=1, keys=['Count',
    return summary1.head(20)

department_name = 'Apparel'
customer_country = 'EE. UU.'
category_names = ["Women's Clothing", "Men's Clothing"]

summary1 = summarize_order_country(delivery_status, department_name, customer_country, category_

fig, ax = plt.subplots(figsize=(5, 4))
summary1.plot.barh(y='Count', alpha=0.8, legend=False, ax=ax)
ax.set_title('Apparel suppliers for EEUU')
ax.set_xlabel('Order Count')
ax.set_ylabel('Order Country')

txt = """The analysis of the data shows that all suppliers selling
clothes in the US through the platform Acme Sports are from Asian
and Oceanic countries. This presents an opportunity for companies
from other regions, especially the UK, to compete in the market by
offering quicker delivery times, as the UK is geographically
closer to the US.

The data reveals that there is room for improvement in the delivery
times of clothing from Asian and Oceanic suppliers, providing a gap
that other suppliers can fill by offering faster delivery times.

By providing customers with faster delivery times, companies from other
```

```
regions could gain a competitive advantage in the US market, leading to
increased market share and profitability."""

plt.text(1, 0.1, txt, fontsize=14, transform=plt.gcf().transFigure)

plt.show()
```



The percentage breakdown of different types of shipping performance. Late delivery is the most common type of shipping issue, accounting for over half of the total at 55.49%. Shipping on time is the next most common type, making up 19.55% of the total. Advance shipping is the third most common type at 21.42%, and shipping canceled is the least common at 3.54%.

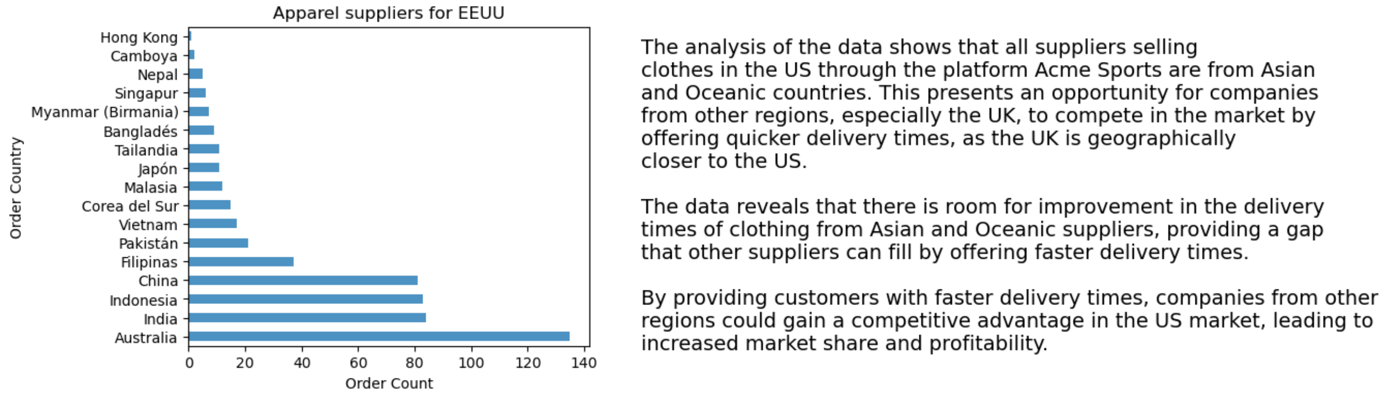


Moreover, if we dig deeper, data reveals that the second class shipping option has the worst delivery record with 74% of shipments arriving late.

Although the standard class has the highest number of orders with advanced shipments, it also has the largest portion of orders arriving late.

The best option appears to be Same Day shipping, with a success rate of 56%, which is the highest among all the available shipping options.

Overall, the data suggests that careful consideration should be given to selecting the appropriate shipping method to ensure that orders are delivered on time and customer satisfaction is maintained.



The analysis of the data shows that all suppliers selling clothes in the US through the platform Acme Sports are from Asian and Oceanic countries. This presents an opportunity for companies from other regions, especially the UK, to compete in the market by offering quicker delivery times, as the UK is geographically closer to the US.

The data reveals that there is room for improvement in the delivery times of clothing from Asian and Oceanic suppliers, providing a gap that other suppliers can fill by offering faster delivery times.

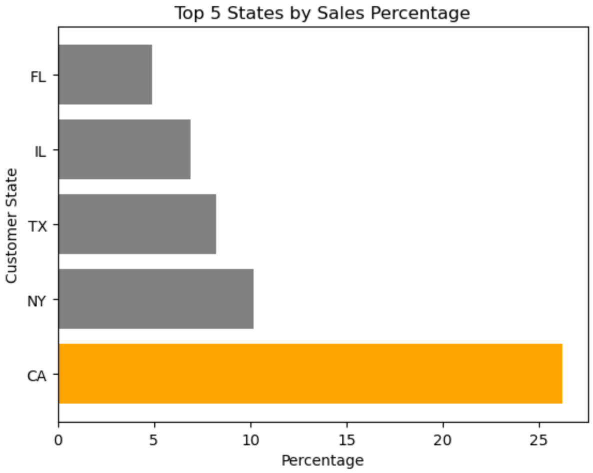
By providing customers with faster delivery times, companies from other regions could gain a competitive advantage in the US market, leading to increased market share and profitability.

Problem 3

In [102…

```
pivot_table = pd.pivot_table(chart_Filtered,
                             index='Customer State',
                             values='Sales',
                             aggfunc='sum')

# add percentage column
pivot_table['Percentage'] = pivot_table['Sales'] / pivot_table['Sales'].sum() * 100

# sort by sales in descending order
pivot_table = pivot_table.sort_values('Sales', ascending=False)
```

```
colors = ['orange', 'grey', 'grey', 'grey', 'grey']
top_states = pivot_table.sort_values(by='Percentage', ascending=False).head(5)
plt.barh(top_states.index, top_states['Percentage'],  color = colors)
plt.title('Top 5 States by Sales Percentage')
plt.xlabel('Percentage')
plt.ylabel('Customer State')

txt = """
Looking at the data, California is clearly the top state in terms of
sales, with a total of $27,893. New York and Texas follow in second
and third place, respectively, with sales of $12,428 and $11,820.

Illinois and Ohio round out the top five, with sales of $7,504 and
$4,512, respectively.

Based on this information, it would be wise to focus sales efforts on
California, as it is clearly the biggest market. However, it's also
worth noting that the other top states, such as New York and Texas,
represent significant sales opportunities as well.

Ultimately, a successful sales strategy will likely involve targeting
multiple states and ensuring that products are tailored to the unique
preferences and needs of each market.




"""

plt.text(1, 0.01, txt, fontsize=14, transform=plt.gcf().transFigure)

plt.show()
```



Looking at the data, California is clearly the top state in terms of sales, with a total of $27,893. New York and Texas follow in second and third place, respectively, with sales of 12, 428*and*11,820.

Illinois and Ohio round out the top five, with sales of $7,504 and $4,512, respectively.

Based on this information, it would be wise to focus sales efforts on California, as it is clearly the biggest market. However, it's also worth noting that the other top states, such as New York and Texas, represent significant sales opportunities as well.

Ultimately, a successful sales strategy will likely involve targeting multiple states and ensuring that products are tailored to the unique preferences and needs of each market.