

Interim Data Analysis Project Report

Group 5

Aqsa Shuja

Iris Engler

Stephy Zhu

**Table of Content**

[Project Overview 1](#_brentieg9hgr)

[Approach to AdventureWorks Data Analysis 2](#_qp3wjj7bpapb)

[Problem 1 - Regional Sales in the Best Performing Country 3](#_fj7p7pgbohfp)

[Problem 2 - Relationship between Annual Leave & Bonus 4](#_zajxrdlkb65f)

[Problem 3 - Relationship between Country and Revenue 5](#_7weuafv5pu5s)

[Problem 4 - Relationship between Sick Leave and Job Title 7](#_2s25cwxuqzil)

[Problem 5 - Relationship between Store Trading Duration and Revenue 9](#_saq32uqpxgf8)

[Problem 6 - Relationship between Size of Stores, Number of Employees and Revenues 10](#_4q7rjrkt80yt)

[Challenges 12](#_1igfv8hu2rm1)

# Project Overview

Adventure Works Cycles is a manufacturer and distributor of bicycles and components in the North American, European and Asian commercial markets, and has over 500 employees.

The objective of this project was to analyse the company’s database in order to explore and understand the relationships between various factors, with a focus on store sales and employee performance. By examining these relationships, we aimed to gain insights that can inform decision-making and potentially optimise business operations.

We used a combination of statistical analysis techniques and data visualisation methods. Correlation analysis assisted us to identify the strength and direction of the relationships between variables. Additionally, descriptive statistics provided insights into the distributions of the data.

# Approach to AdventureWorks Data Analysis

The AdventureWorks dataset is a large dataset with 5 schemas and over 20 tables. In this project we used tables under two schemas - HumanResources and Sales. The dataset contains built-in views as well, which were also used in this project.

Tables and Views used in this project：

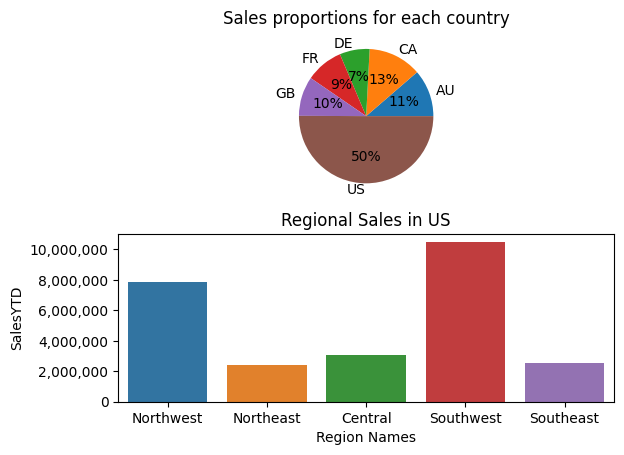
* Sales.SalesTerritory
* Sales.SalesPerson
* Sales.vStoreWithDemographics
* Sales.vStoreWithAddresses
* HumanResources.Employee
* HumanResources.EmployeePayHistory

Tools used in this projects are as follows:

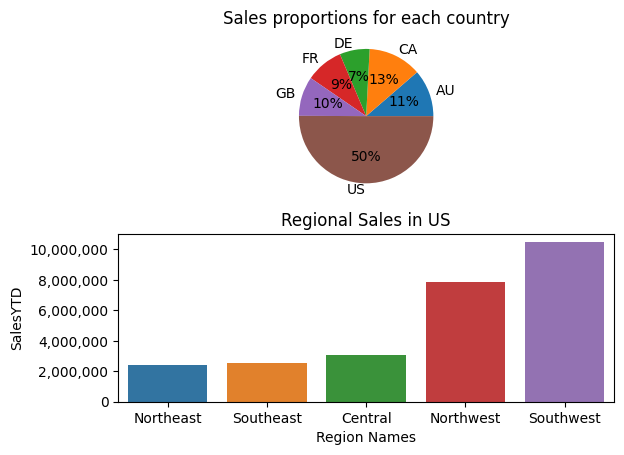
* **SQL Server Management Studio** - for first step data extraction, we used SQL Server Management Studio and SQL queries to select necessary columns from SQL data and export as .csv files for further analysis.
* **Python and Jupyter** - we used Python to perform data analysis and visualisation. Jupyter was our main platform for running Python codes.
* **Libraries** - three python libraries were used in this project. Pandas for loading datasets and performing data analysis, Matplotlib.pyplot and Seaborn for Data Visualisation.

# Problem 1 - Regional Sales in the Best Performing Country

By extracting data from Sales.SalesTerritory table, we retrieved a table that contains Sales Year to Date for every region and country accordingly. By plotting a pie chart of sales proportions in countries, we can clearly see that the United States is taking the lead, by contributing half of the sales for the company.



After finding out that the United States is the best performing country, we can take a close look at regional sales in the United States. The bar chart below shows the sales for each region in America. The Southwest region is the top performing region with over 10 Million dollars sales, followed by the Northwest region. All other three regions performed similarly, within the range of 2-3 million.



# 

# Problem 2 - Relationship between Annual Leave & Bonus

Do people get a higher bonus if they take less annual leave?

The answer can be found in the scatter plot. This scatter plot shows the distribution of employee bonuses and the remaining annual leave hours. The distribution in the plot is quite random, which suggests that bonuses are not related to annual leave.

# 

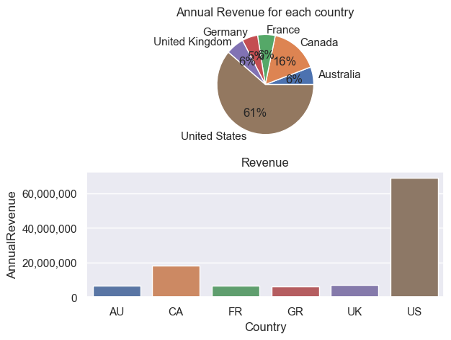
A correlation test on the two factors has also been run, and the result is 0.38. This result indicates a weak positive relationship.

In conclusion, while taking less annual leave might result in a higher bonus, the relationship is very weak. There may still be other factors that could be affecting bonuses as well.

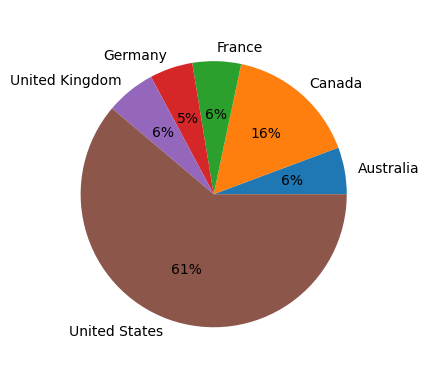
# 

# Problem 3 - Relationship between Country and Revenue

We generated the following visualizations to plot revenue generated by each country:



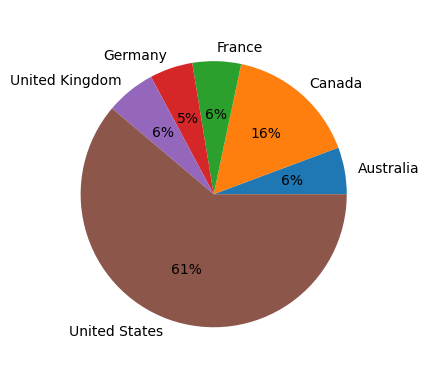
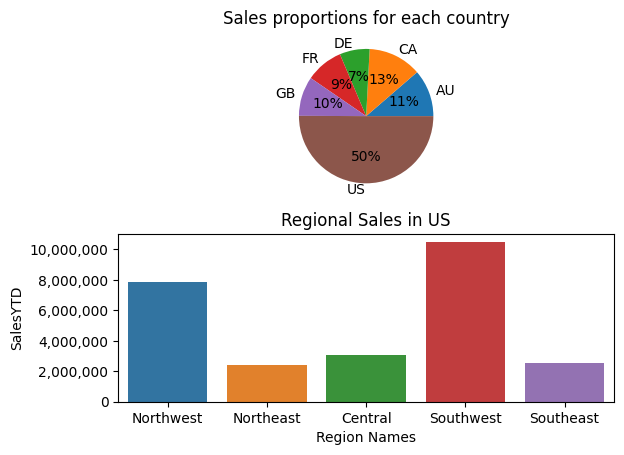
In the Bar Plot, The x-axis represents the “Country” and the y-axis represents the “Annual Revenue” values. Bars represent revenue generated by each country.



Pie Plot shows the percentages of Annual Revenue generated by each country.

We can see that the United States produced significantly higher revenue compared to other countries. Canada holds the second place. The rest of the countries have similar performance in terms of generating revenue for the company.

When we compare the revenue percentages to the sales percentages, we can see that for the United States and Canada the revenue is higher than the sales percentage which means that these countries generate revenue by sales of products plus other sources. Hence, these countries are more efficient and have a higher percentage of revenue.

Left chart: Pie Plot of percentages of Revenue by country

Right chart: Pie Plot of percentages of revenue by country

However, Other countries have revenue percentages lower than sales percentages. This shows that there are factors that negatively impact the country’s ability to make revenue for the company.

From this analysis we can conclude that the company performs best in terms of revenue in the local market within the United States.

# Problem 4 - Relationship between Sick Leave and Job Title

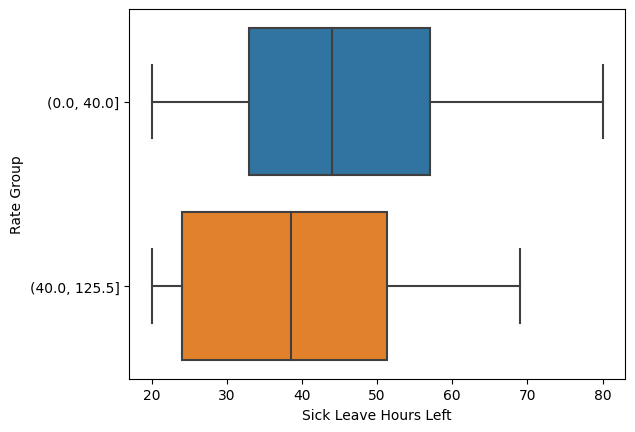
We generated a scatter plot to show Sick leave hours available for the 67 unique job titles.

The x-axis represents “Average sick leave hours left” and the y-axis represents “Job Titles”. Each point on the plot represents the average sick leave hours available for every Job Title. The plot does not show any relationship.



However, If we associate a quantitative value to Job Titles we can establish a relationship.

We generated the following box plot to explain the relationship between different Job Titles and Sick Leave Hours left when we categorize Job titles based on pay rates.



In the Box Plot, we have “Sick Leave Hours Left” on the x-axis and “Pay Rate Groups” on the y-axis.

We divided pay rates into the following two groups:

* $0 - $40 (low pay rate group)
* $40 - $125 (high pay rate group)

​​The median sick leave hours left for the low pay rate group is higher than the high pay rate group. We can conclude that the low pay group takes less sick leave as compared to the high pay rate group.

# Problem 5 - Relationship between Store Trading Duration and Revenue

Trading Duration refers to the number of years the store has been operating and was calculated using the ‘YearOpened’ field.

We generated the following box plot that displays the distribution of the annual revenue of stores based on their years since opening.

The x-axis represents the quartile based categories of “Years Since Opening” and the Y-axis represents the “Annual Revenue” values.



We can see that the oldest stores (Q4) have significantly lower revenues than the newer (Q1-Q3) stores. The median revenue for the newer stores is the same as the top quartile of revenues for the older stores.

Other things to note are that even though the medians are the same for stores in Q1-Q3, the worst performing stores in Q1 do better than the worst performing stores in Q2 and Q3 (the newest stores have the smallest inter quartile range).

Although we can conclude that the oldest stores have significantly lower revenues, this does not necessarily mean that these stores should be closed. Other factors such as operating costs (eg. older stores could be smaller sized stores) and brand loyalty (older stores could have a loyal customer base) should be taken into account.

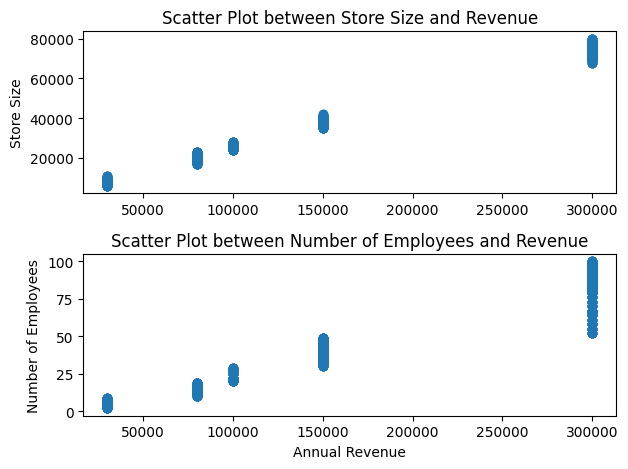
# Problem 6 - Relationship between Size of Stores, Number of Employees and Revenues

We generated the following visualizations:

A heatmap plot that visually represents the correlation matrix between the store size, number of employees and annual revenue:



Two scatter plots, the first showing the relationship between store size and annual revenue and the second showing the relationship between the number of employees and annual revenue.



We can see from the scatter plots that there don’t seem to be many medium sized stores. Around 70% of the stores are under 50000 square feet and the rest are large stores.

All visualizations show that there is a very strong positive relationship between annual revenue and both the size of the store and the number of employees.

The relationships are expected ones. Larger stores typically have a larger physical space and customer base and require more employees. They also have a wider range and quantity of products compared to smaller stores and will therefore generate more revenue.

# Challenges

* Installing the SQL Server Management Studio and connecting between it and the SQL server was not simple and was time consuming. One of our team members had the added challenge of installing the software on a MAC computer. This challenge was overcome by following the steps detailed in guides such as YouTube videos.
* Understanding the data contained in some of the fields was not always straightforward. For example, with both the ‘Sick Leave’ and ‘Annual Leave’ fields we weren’t sure whether the values indicate the number of available or the number of used hours. We sourced the data dictionary and used the descriptions for these fields to solve the problem.
* The last challenge related to the exploration of data in order to determine which fields should be used to investigate relationships. Some of the fields required further information. For example, we weren’t sure whether the Annual Revenue field could be used as it contains repeated values and only whole numbers. We approached the client and they confirmed that it is a valid field.

# 