A dive into AdventureWorks dataset

Setting the Scene

This is a comprehensive deep dive into Adventure Works company data with the following questions central to the analysis:

1. What are the regional sales in the best performing country?
2. What is the relationship between annual leave taken and bonus?
3. What is the relationship between Country and Revenue?
4. What is the relationship between sick leave and Job Title (PersonType)?
5. What is the relationship between store trading duration and revenue?
6. What is the relationship between the size of the stores, number of employees and revenue?

**Overview of the dataset:**

Company name: Adventure Works

Company type: Manufacturer of bikes and biking equipment

Sales channels:

* Business to business (retail stores)
* Online sales direct to customer

**Overview of data:**

Period data taken from: 31/5/11 to 30/6/14

Staff data:

* Total staff: 290
* 5 Pay grade tiers (inc CEO)
* 67 titles
* 3 shift patterns
* 14 bonus related staff

Sales data:

* 121,317 item sales
* $109,846,381 revenue (over all data)
* 19,972 separate customer accounts
* 701 retail outlets distributed to

Trading regions: 6

Territories operating in: 10

Multi language culture – 7 languages + invariant country

# Question 1 – Regional Sales in the Best Performing Country

A pie chart with different colored circles

Description automatically generatedA pie chart with different colored sections

Description automatically generatedUpon querying the sales.SalesTerritory table, grouping by CountryRegionCode and ordering by SalesLastYear in descending order (see appendix item I), it was discerned that the United States stands out as the best-performing country based on raw sales numbers. Subsequently, the U.S. sales data was disaggregated by region, extracting region (Name), CountryRegionCode, SalesYTD, and SalesLastYear from the same table (see appendix item II), and grouped by region to obtain insights for the Northwest, Northeast, Central, Southwest, and Southeast regions. The data was exported to a .csv file and imported into a pandas dataframe. Two pie charts were constructed, delineating the breakdown of sales in the U.S., with one reflecting year-to-date sales and the other focusing on the preceding year.

(Code for pie charts in appendix item III)

The emphasis on SalesLastYear was motivated by the consideration of potential biases and incomplete data in SalesYTD, such as seasonal fluctuations. Analysis revealed that the Southwest emerged as the highest-performing region, followed by the Southeast, Northeast, Northwest, and Central U.S. Given the data year to date only included data up to June, it is likely we can conclude that the western side of the US drives sales for AdventureWorks in the first half of the year due to weather conditions.

**Next steps**

Investigate further the reasons for change in regional performance in comparison to previous year.

Investigate further the drivers as to why certain regions outperform others, e.g. buying behaviours, online sales, relationships with stores and regional population size vs sales.

# Question 2 – Relationship Between Annual Leave Taken and Bonus

A meticulous examination of the Human Resource schema identified 14 staff members (see appendix item IV), all sales representatives, participating in a bonus scheme. Notably, a correlation between bonus and vacation hours was explored through a scatter plot, revealing an absence of a clear correlation between the two variables, as seen in the scatter plot below. However, with such a small sample size it is unlikely the findings can be considered robust.

A graph with blue dots

Description automatically generated

(see appendix item V for scatter plot code)

**Summary**

* There is no clear correlation between the variables so no recommendations to make changes to the Sales Representatives annual leave to influence bonus.

**Next steps**

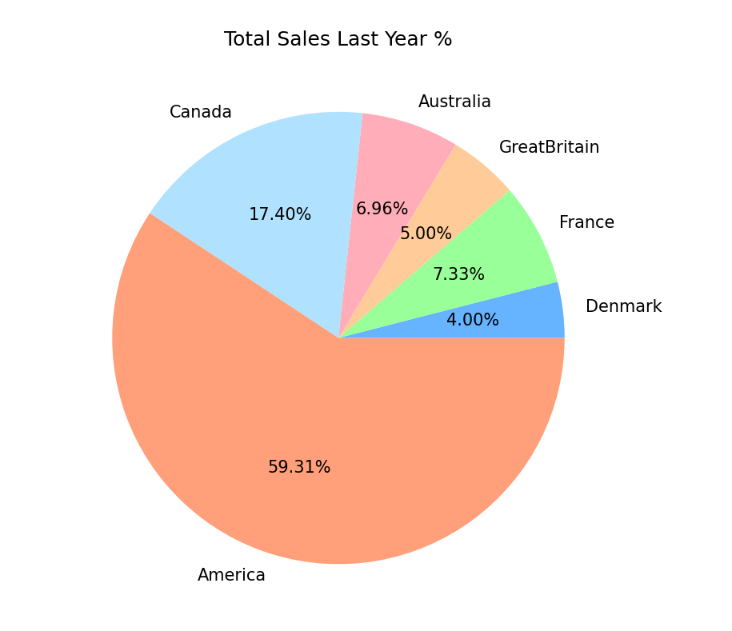
* Due to sample size and robustness of data check industry best practices to see if there are any correlations between bonus and annual leave.

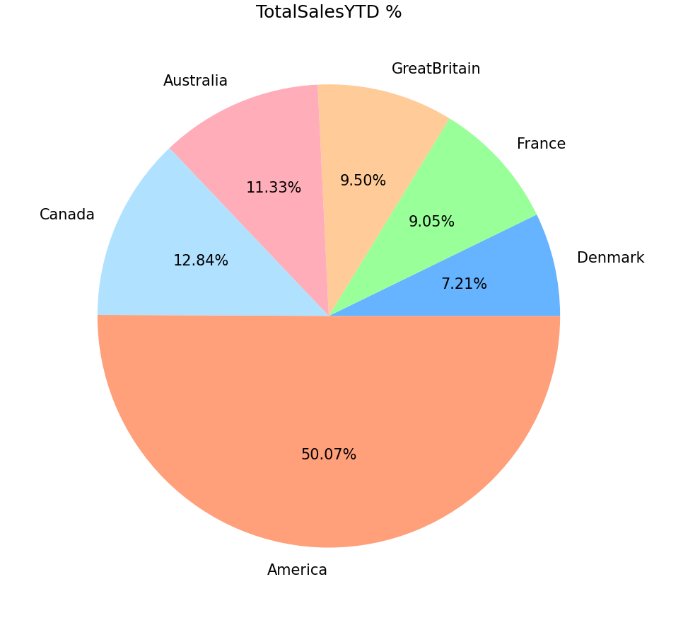
# Question 3 – Relationship Between Country and Revenue

To address this query, referring to the Sales.SalesTerritory table in SQL, extracting SalesLastYear, CountryRegionCode, and SalesYTD columns (see appendix item I). Total sales by country were calculated using SUM and GROUP BY functions, resulting in TotalSalesLastYear and TotalSalesYTD. Subsequently, a SQL view was created, and the data was exported to a CSV file, which was then imported into a pandas dataframe for visualization. Matplotlib, pandas, numpy, and seaborn libraries were employed to create a horizontal bar graph showcasing the disparity in sales between the last year and the current year to date.

A graph of sales and sales

Description automatically generated

Additionally, pie charts were generated for sales last year and sales this year, providing further insights.

(see appendix item VI for code for visualisations in Q3)

**Findings**

* America represents just over 50% of the sales year to date with marginal differences in the other 5 country zones.
* All areas have grown since last year.
* Australia, Great Britain, France and Denmark are noted as the areas of most significant growth year on year.
* Whilst America and Canada also showed growth, this was significantly less than the other 4 regions.

**Next steps**

* Investigate drivers for change in year on year and between countries performance:
  + Internally
    - Distribution channels (store/online)
    - Product splits/buying patterns
    - Store relationships
  + Externally
    - Market saturation
    - Economic factors
    - Buying behaviours

# Question 4 – Relationship Between Sick Leave and Job Title

Analysis of the Human.Resources employee table involved grouping by job title and ordering by the average sick leave hours (see appendix item VII). After converting this query result to a pandas dataframe, the following graph was created:

A red and white gradient

Description automatically generated

(see appendix item VIII for chart code)

As seen from the chart above, this provides an overview by department, however as a lot of job titles have a minimal sample size, the data cannot be seen as robust when taken to this level and may be easily affected by small deviations.

The report therefore investigates the relationship of sick leave by department, hierarchy level, shift work and gender.

*By department*

By joining the HumanResources.Derpartment, HumanResources.EmployeeDepartmentHistory and HumanResources.Employee tables (as shown in appendix item IX). Avg vacation hours were also included to see if there was a correlation with sick leave hours, which there was, suggesting that some departments have roles that are more tiring than others and thus employees are more likely to take time off work. The results were then sorted by the average sick leave hours and are presented in the chart below:

A graph of sick hours

Description automatically generated

(see appendix item X for chart code)

As you can see, the departments with roles most likely to take more sick leave hours include ‘shipping and receiving’ and ‘facilities and maintenance’, those less likely to have tiring roles, are ‘sales’ and ‘engineering’ taking fewer sick hours, likely due to being interesting and non-labour-intensive.

*By hierarchy*

Job titles summarised by their level taken from the HumanResources.Employee table in the dataset has a field “OrganizationLevel” which summarizes each role’s standing in the company hierarchy (see appendix item XI). The roles were grouped by this field the average sick leave hours for each bracket are plotted in the chart below:

A graph of average sick leave hours for each hierarchy level

Description automatically generatedAs the chart displays, the CEO took the most average sick leave hours last year, however it’s worth noting there is only one sample and data can be skewed by small changes. Tier 1 has the least sick leave (mainly upper management) whilst results were fairly consistent between the others.

(code for bar chart in appendix item XII)

*A graph of blue rectangular shapes

Description automatically generatedBy shift pattern*

After grouping by shift type (see appendix item XIII), it was identified that although staff on the day and evening shift patterns both take on average 44 hours sick leave per year, staff on night shifts take 48 hours on average. They also take more vacation hours, showing that working the night shift is a tiring job and they feel the need to take more time off work.

(see appendix item XIV for bar chart code)

*By gender*

Employees were also grouped by gender to investigate for any potential trend but none was observed, with both genders taking on average 45 sick leave hours per annum, the same as the company overall average (see appendix item XV).

**Findings**

* Sick leave by department was most prevalent amongst ‘shipping and receiving’ and ‘facilities and maintenance’. The assumption that these are more manual roles and staff may be more likely to be injured/tired from the work environment.
* Sick leave was least amongst ‘sales’ and ‘engineering’ with the assumption that these are least manual.
* Sick leave by hierarchy was least prevalent at the tier 1 level (mainly upper management), with consistency between tiers 2-4. The assumption that upper-level management are least likely to take off time to focus on work even if worn down/ill.
* By shift pattern, as expected night workers took more sick leave.
* There was no difference in sick leave patterns between gender.

**Next steps**

* Check company sick leave patterns against industry standard.
* Investigate departments with higher levels of sick leave for reasons of leave. If down to tiredness/worn down maybe review practices, e.g., more break periods, free health checks etc.
* Investigate best practices for night shift work to understand is sick leave can be reduced.

# Question 5 – Relationship Between Store Trading Duration and Revenue

Utilizing the sales.Store table, information was extracted on each supplied store, with special attention to the demographics field containing hyperlinks, necessitating the use of the Sales.vStoreWithDemographics view. The YearOpened field and the average of the AnnualRevenue field were selected, grouped by YearOpened (see appendix item XVI) to illustrate the average annual revenue for each opening year across all stores.

Post-exporting the results to a .csv file and importing into pandas, a store age in years column was introduced, calculated by subtracting the opening year from 2014. A scatterplot was then generated with trading duration on the x-axis and average revenue on the y-axis, accompanied by a line of best fit. The analysis indicated a slight negative correlation between store age and average revenue, supported by a calculated correlation of -0.22 (rounded to 2 d.p.), signifying a weak negative correlation.

A graph with purple dots and a yellow line

Description automatically generated

(code in appendix item XVII)

**Findings**

* There is a weak negative correlation between the relationship of store trading duration and revenue. This needs investigating further to understand if there are any specific drivers influencing this relationship.

**Next steps**

* Interview sales representatives to understand relationship with stores.
* Understand percentage of store sales AdventureWorks account for and if share has grown/decreased with store age.
* Review market at regional level to understand competition within stores.
* Identify stores target audience, e.g., mainstream/niche product range.

# Question 6 - What is the relationship between the size of the stores, number of employees and revenue?

Utilising the same view as in question 5, unpacking the sales.Store table, the 701 stores supplied by the company were put into brackets, classifying them by store size in square feet by using a CASE WHEN statement to assign them into bins for each 10k square feet (see appendix item XVIII). The average revenue for each bracket was also selected and plotted in a bar chart.

A graph of a growing bar

Description automatically generated with medium confidenceAs the chart shows, there appears to be a correlation between size of store and average revenue, which makes sense given larger stores are likely to serve more customers, stock more products, shift more high-ticket items, etc.

(code for chart in appendix item XIX)

*Correlation by store size, number of employees and revenue.*

The following 3 tables using the data from the sales.store table, found there is almost exactly the same correlation between ‘Revenue vs Number of Employees’, ‘Revenue vs Square Feet’ and ‘No of Employees vs Square Feet’ which backs up the findings above.

A graph with blue dots and numbers

Description automatically generatedA graph with blue dots

Description automatically generatedA graph of a graph showing a number of blue squares

Description automatically generated with medium confidence

**Findings**

* There is an almost direct correlation between store size, number of employees and revenue, however at each point the data seems to plateau.

**Next steps**

* Investigate relationships with store:
  + is there potential to take a larger share of store sales by running promotion either on sales to stores, or end user.
* Investigate why data is congregated around certain points on the 3 scatter plots.

Appendix:

Item I:

A screenshot of a computer

Description automatically generated

Item II:

A screenshot of a computer

Description automatically generated

Item III:

A screen shot of a computer program

Description automatically generated

A screen shot of a computer program

Description automatically generated

Item IV:

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

Item V:

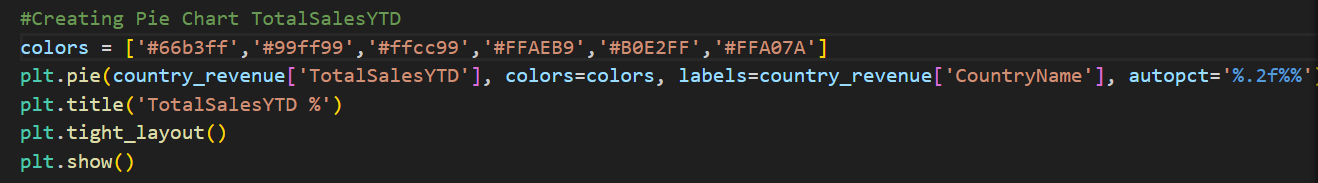
A computer screen with colorful text

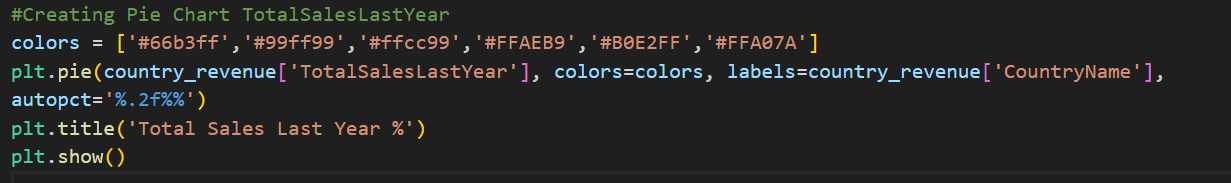
Description automatically generated

Item VI:

A screenshot of a computer code

Description automatically generated





Item VII:

A screenshot of a computer

Description automatically generated

Item VIII:

A screenshot of a computer program

Description automatically generated

Item IX:

A screenshot of a computer

Description automatically generated

Item X:

A screen shot of a computer program

Description automatically generated

Item XI:

A screenshot of a computer

Description automatically generated

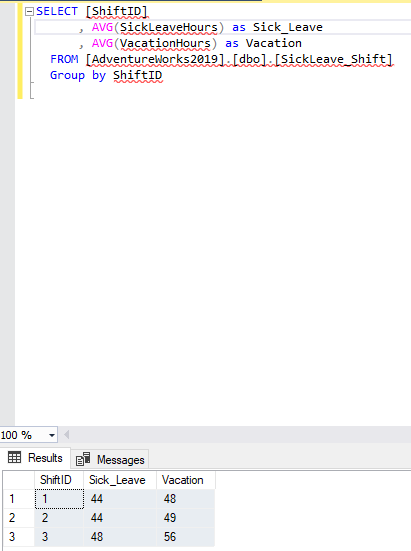
Item XII:

A computer screen shot of text

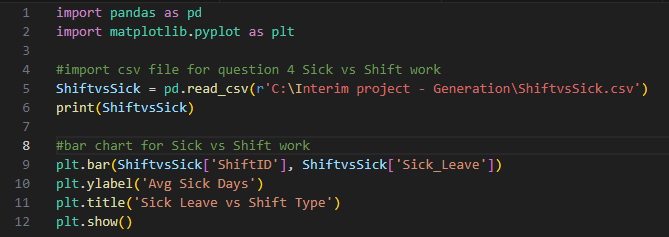
Description automatically generated

Item XIII:

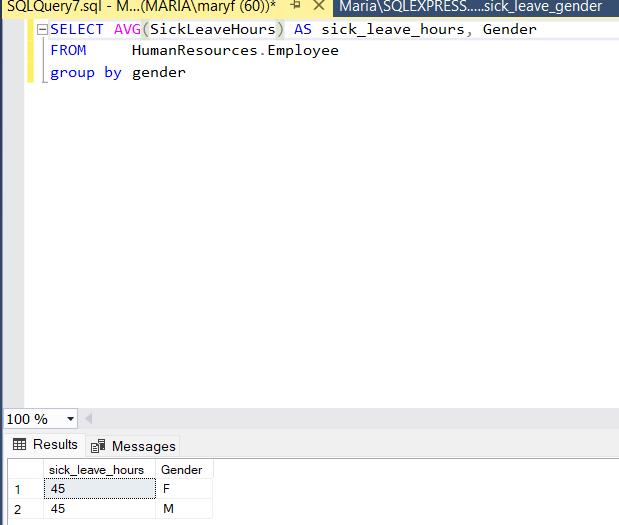
A screenshot of a computer

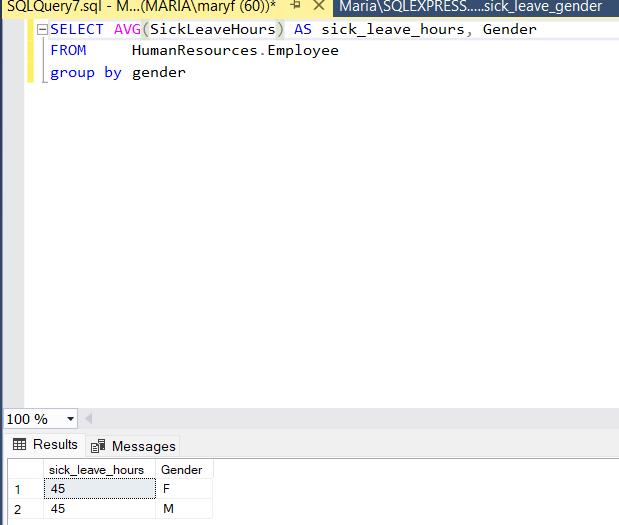
Description automatically generated

Item XIV:



Item XV:





Item XVI:

A screenshot of a computer

Description automatically generated

Item XVII:

A computer code on a black background

Description automatically generated

Item XVIII:

A screenshot of a computer

Description automatically generated

Item XIX:

A black background with colorful text

Description automatically generated