Colorado Motor Vehicles Sales Analysis

Project Report

Submitted to

UNIFIED MENTOR

Submitted by

Mr. Abdul Razzaq

Acknowledgement

I would like to express my sincere gratitude to my faculty guide for their continuous guidance, valuable suggestions, and encouragement throughout the completion of this project. I am also thankful to my institution for providing the necessary resources and learning environment to carry out this work. Finally, I would like to acknowledge the open-source R community for the libraries and tools that made data analysis and model development possible.

Table of Contents

Chapter 1

Abstract

This project analyses motor vehicle sales data across Colorado counties (2008–2015) to identify sales trends, regional performance, and seasonality.

Using R for exploratory data analysis and ARIMA-based forecasting, and Power BI for visualization, the study finds steady growth after 2010, seasonal peaks in Q3–Q4, and Arapahoe and El Paso as the top-performing counties.

Objective

The objective of this analysis is to understand sales behaviour over time and across regions to identify key factors influencing performance and forecast future sales trends. We aim to:

- Analyse historical sales patterns across different counties.
- Evaluate seasonal and quarterly variations in sales.
- Using R to forecast statewide vehicle sales for future quarters.
- Develop a visual dashboard for interactive insights using Power BI.

Chapter 2

Dataset Collection

The given dataset provided in the pdf is: Colorado Sales Data

Dataset contains the following:

- Year: The calendar year in which the sales data was recorded.
- Quarter: The quarter of the year during which the sales were made. The quarters are divided as follows:
  - Q1: January to March
  - Q2: April to June
  - Q3: July to September
  - Q4: October to December
- County: The name of the county in Colorado where the sales were recorded.
- Sales: The total dollar amount of motor vehicle sales in the specified county and quarter.

Chapter 3

Tools and technologies

R Studio:

R was used as the primary environment for data cleaning, wrangling, exploratory analysis, and forecasting because of its strong statistical capabilities and extensive libraries.

Key R Packages Used:

tidyverse: For data manipulation (dplyr), reshaping (tidyr), and data import (readr).

Tasks: Cleaning raw data, handling missing values, creating derived variables.

ggplot2: For creating high-quality visualizations such as line charts, bar charts, and heatmaps.

Tasks: EDA plots showing trends, seasonality, county performance.

forecast (ARIMA): For time series modelling and forecasting future sales.

lubridate: For handling date conversions from Year/Quarter format.

R was chosen because it is ideal for statistical analysis, reproducible research, and quickly generating analytical outputs for BI tools**.**

Microsoft Power BI:

Power BI was used to create an **interactive dashboard** that presents insights in a dynamic and user-friendly manner.

Power BI Components Used in this project**:**

Power Query Editor:

Used for data transformation, adding calculated columns (like County Name), and building a custom Date field for time intelligence.

Data Model (Relationships & DAX):

- Created a Date Table, defined relationships, and wrote DAX measures for:
  - Total Sales
  - Highest Quarterly County Sales
  - Running totals and averages

Visualizations:

- Line charts (Yearly trends)
- Bar charts (County comparisons, Quarterly sales)
- Filled Map (County-wise geographic analysis)
- Table and Matrix visuals
- Slicers (Year, County)

Interactivity Features:

- Tooltip page

Power BI was chosen because it enables interactive data exploration and is widely used in corporate analytics environments.

Microsoft Excel:

Excel was used for initial data inspection, quick value checks and verifying CSV formats before loading into R and Power BI.

Chapter 4

Data Preparation

- Downloaded the data from the provided link, i.e., Colorado_motor_vehicles_sales in csv format.
- Import the data into R as:

```
df <- read.csv("C:/Users/abdul/Downloads/colorado_motor_vehicle_sales.csv")
```

- Preview the data:

```
head(df)
str(df)
summary(df)
colSums(is.na(df))
length(unique(df$county))
df %>% count(county)
```

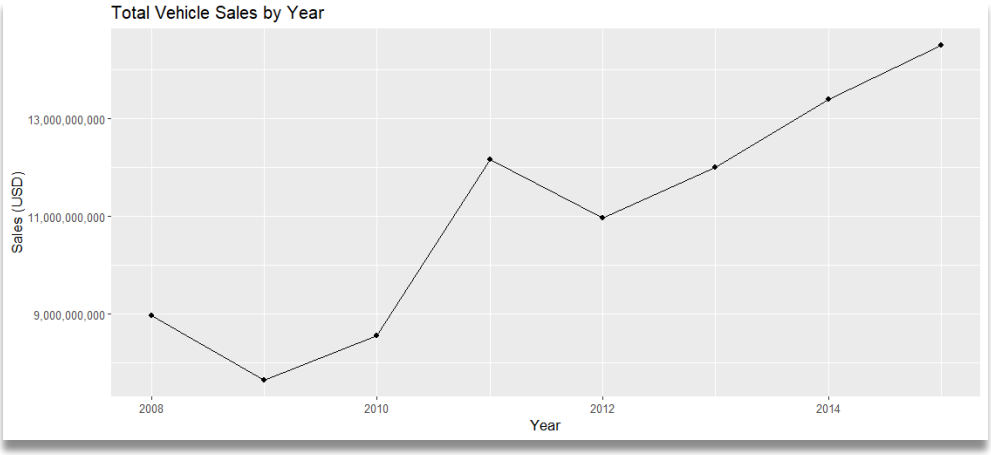
```
> df %>% count(county)
                county  n
1                Adams 32
2            Arapahoe 32
3              Boulder 28
4  Boulder/Broomfield  4
5           Broomfield 28
6               Denver 32
7              Douglas 32
8              El Paso 32
9              Fremont 32
10            Garfield 32
11           Jefferson 32
12            La Plata 32
13             Larimer 32
14                Mesa 32
15              Pueblo 32
16      Rest of State 25
17                Weld 32
```

```
> head(df)
  year quarter             county     sales
1 2008       1              Adams 231609000
2 2008       1           Arapahoe 550378000
3 2008       1 Boulder/Broomfield 176771000
4 2008       1             Denver 200103000
5 2008       1            Douglas  93259000
6 2008       1            El Paso 325737000
> str(df)
'data.frame':   501 obs. of  4 variables:
 $ year   : int  2008 2008 2008 2008 2008 2008 2008 2008 2008 2008 ...
 $ quarter: int  1 1 1 1 1 1 1 1 1 1 ...
 $ county : chr  "Adams" "Arapahoe" "Boulder/Broomfield" "Denver" ...
 $ sales  : int  231609000 550378000 176771000 200103000 93259000 325737000 9911000 55701000 277611000 19558000 ...
> summary(df)
      year         quarter         county             sales
 Min.   :2008   Min.   :1.000   Length:501         Min.   :  6274000
 1st Qu.:2010   1st Qu.:2.000   Class :character   1st Qu.: 61482000
 Median :2012   Median :3.000   Mode  :character   Median :138582000
 Mean   :2012   Mean   :2.503                      Mean   :176058483
 3rd Qu.:2014   3rd Qu.:4.000                      3rd Qu.:224158000
 Max.   :2015   Max.   :4.000                      Max.   :916910000
> colSums(is.na(df))
   year quarter  county   sales
      0       0       0       0
> length(unique(df$county))
[1] 17
```

Exploratory Data Analysis

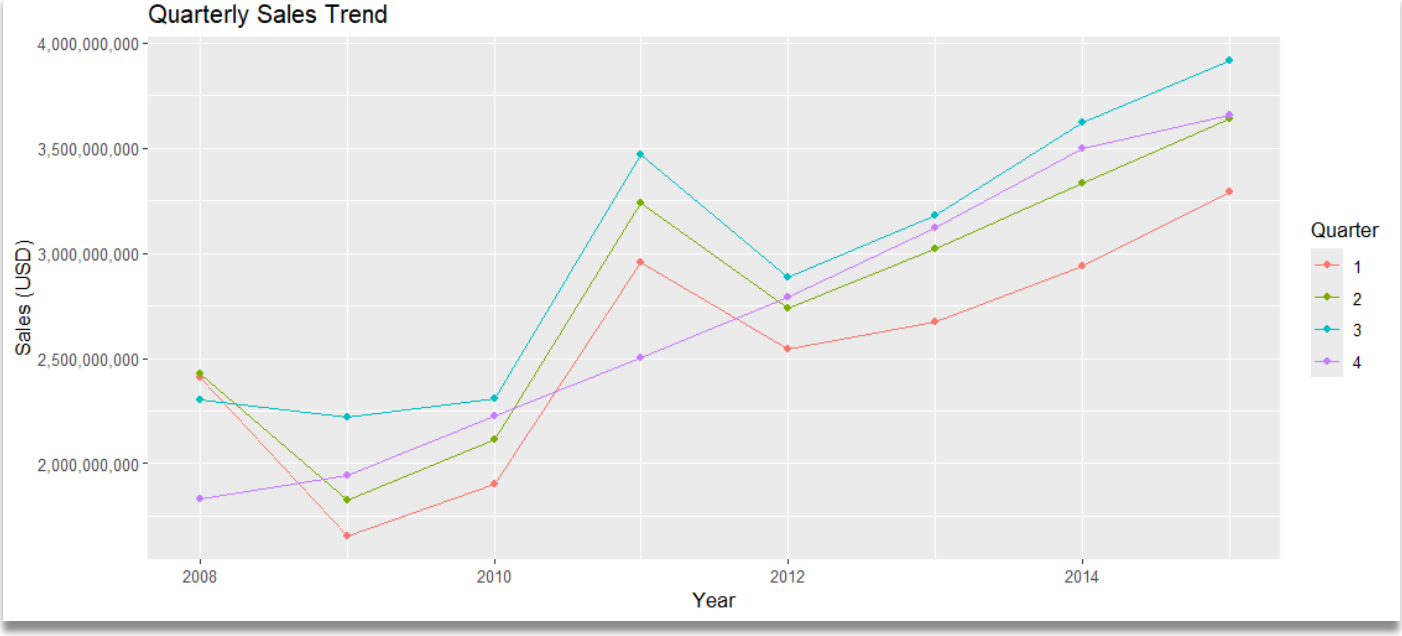
- Plot Yearly sales trend:

```
#Yearly Sales Trend

yearly_sales <- df %>%
  group_by(year) %>%
  summarise(total_sales = sum(sales))

ggplot(yearly_sales, aes(year, total_sales)) +
  geom_line() + geom_point() +
  labs(title="Total Vehicle Sales by Year",
       x="Year", y="Sales (USD)") +
  scale_y_continuous(labels = comma)
```



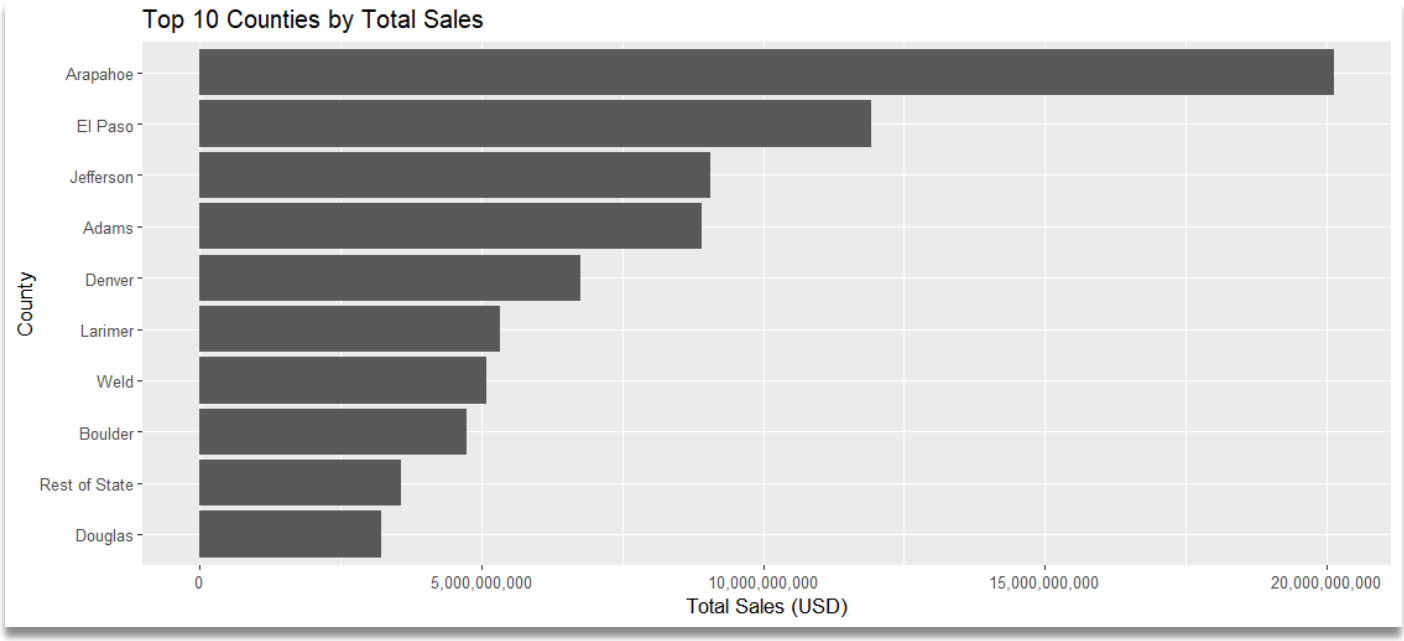
- Plot quarterly sales trends:

```
#Quatarly Trend

quarterly_sales <- df %>%
  group_by(year, quarter) %>%
  summarise(total_sales = sum(sales))

ggplot(quarterly_sales,
       aes(year, total_sales, color=factor(quarter))) +
  geom_line() + geom_point() +
  labs(title="Quarterly Sales Trend",
       color="Quarter", x="Year", y="Sales (USD)") +
  scale_y_continuous(labels = comma)
```



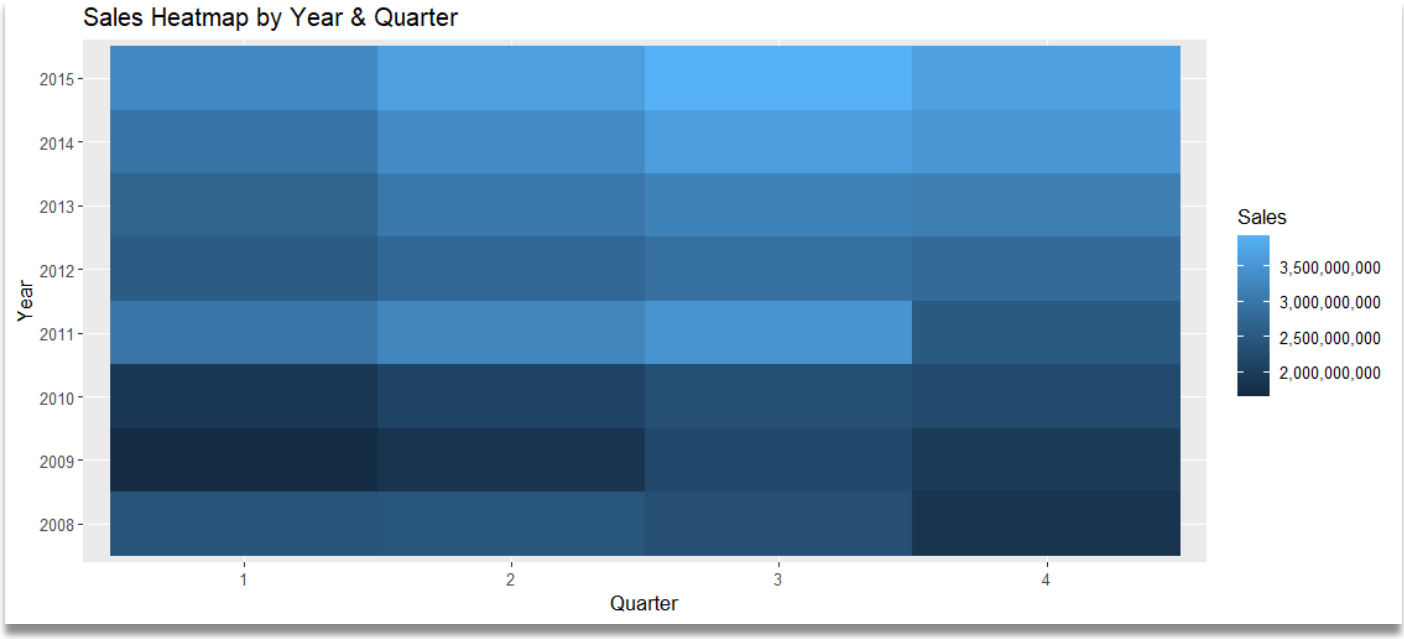
- Plot top 10 counties by sales:

```
#Top 10 counties

top_counties <- df %>%
  group_by(county) %>%
  summarise(total_sales = sum(sales)) %>%
  arrange(desc(total_sales)) %>%
  slice(1:10)

ggplot(top_counties, aes(reorder(county, total_sales), total_sales)) +
  geom_col() +
  coord_flip() +
  labs(title="Top 10 Counties by Total Sales",
       x="County", y="Total Sales (USD)") +
  scale_y_continuous(labels = comma)
```



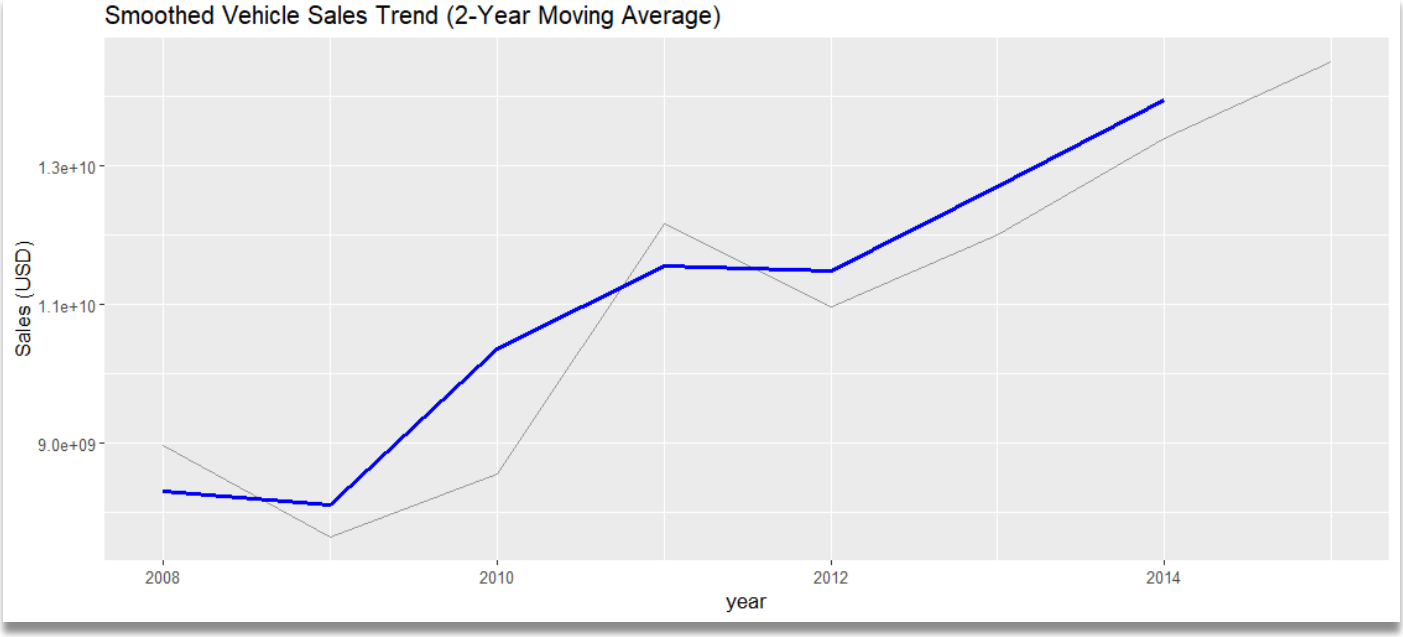
- Plot heatmap of total sales by counties over the years in different quarters:

```
#Quarter heatmap

df %>%
  group_by(year, quarter) %>%
  summarise(total_sales = sum(sales)) %>%
  ggplot(aes(factor(quarter), factor(year), fill=total_sales)) +
  geom_tile() +
  scale_fill_continuous(labels = comma) +
  labs(title="Sales Heatmap by Year & Quarter",
       x="Quarter", y="Year", fill="Sales")
```



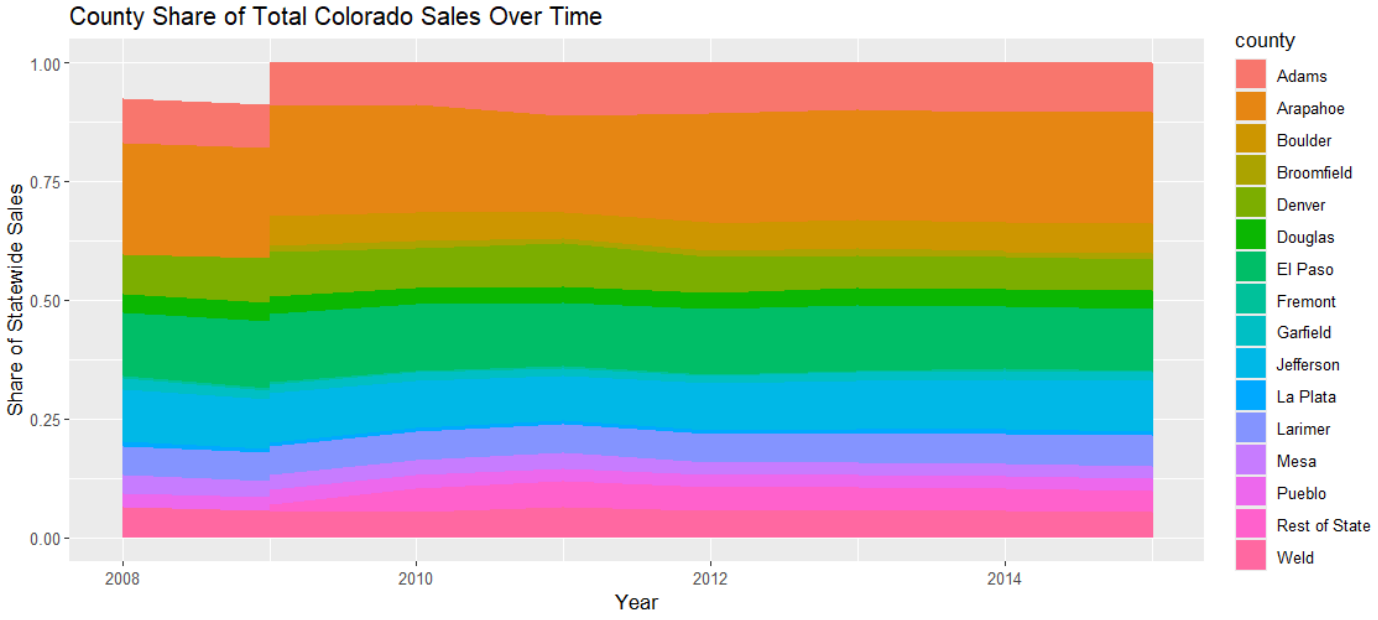
- Plot trend smoothing (running average):

```
#Trend Smoothing

library(zoo)
yearly_sales$moving_avg <- rollmean(yearly_sales$total_sales, 2, fill = NA)

ggplot(yearly_sales, aes(x = year)) +
  geom_line(aes(y = total_sales), color = "grey60") +
  geom_line(aes(y = moving_avg), color = "blue", size = 1.2) +
  labs(title = "Smoothed Vehicle Sales Trend (2-Year Moving Average)",
       y = "Sales (USD)")
```



- Plot County Share of Total Colorado Sales Over Time:

```
#County Share Over Time

df %>%
  group_by(year, county) %>%
  summarise(total_sales = sum(sales)) %>%
  group_by(year) %>%
  mutate(share = total_sales / sum(total_sales)) %>%
  ggplot(aes(x = year, y = share, fill = county)) +
  geom_area() +
  labs(title = "County Share of Total Colorado Sales Over Time",
       y = "Share of Statewide Sales", x = "Year")
```



• Statistical Analysis and Forecast

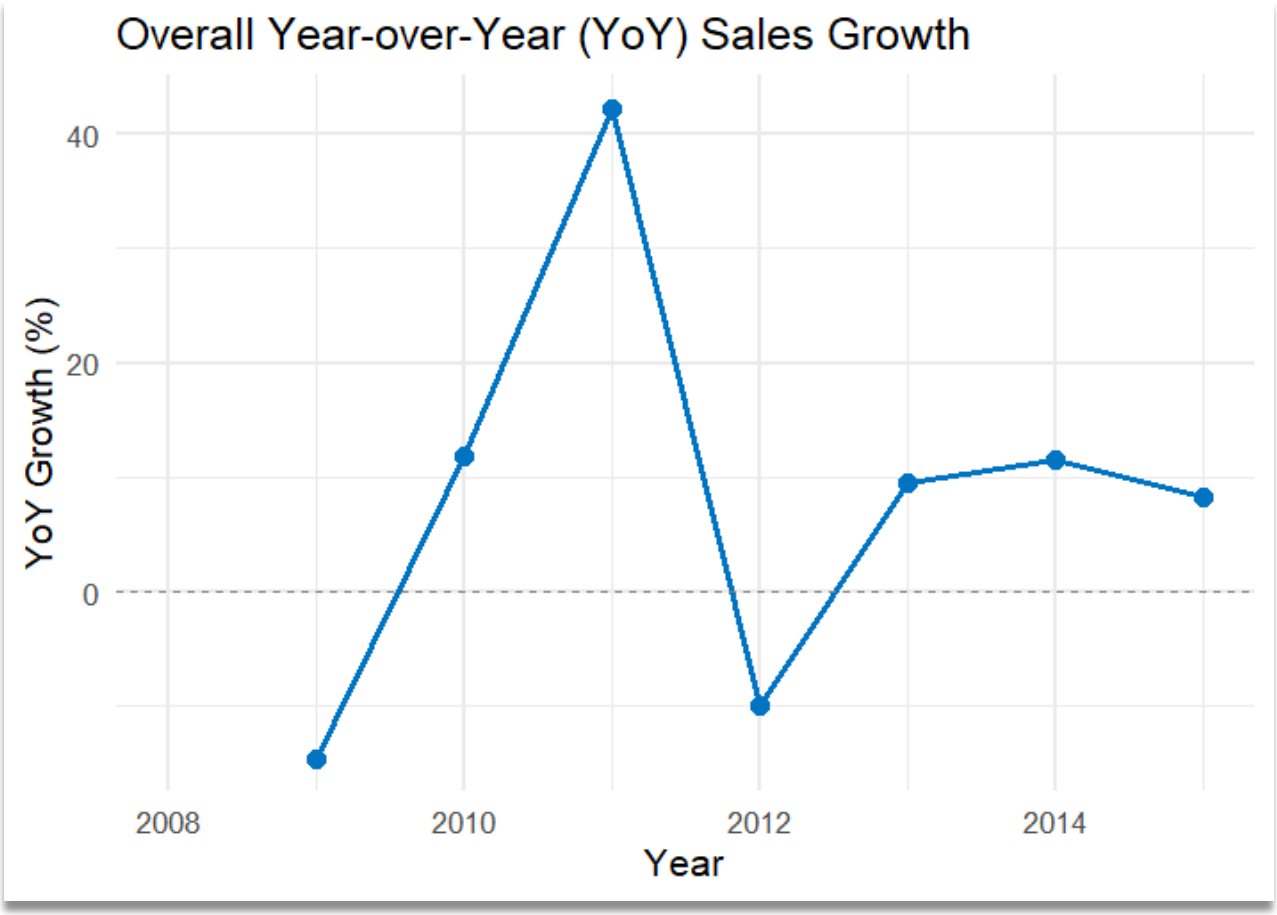
➢ Year over year growth in sales.

```
#Year over Year growth

yoy <- yearly_sales %>%
  mutate(yoy_growth = (total_sales - lag(total_sales))/lag(total_sales))

yoy
```

year	total_sales	yoy_growth
<int>	<dbl>	<dbl>
2008	8965561000	NA
2009	7652500000	-0.146
2010	8556088000	0.118
2011	12170441000	0.422
2012	10960876000	-0.0994
2013	12000615000	0.0949
2014	13392487000	0.116
2015	14506732000	0.0832

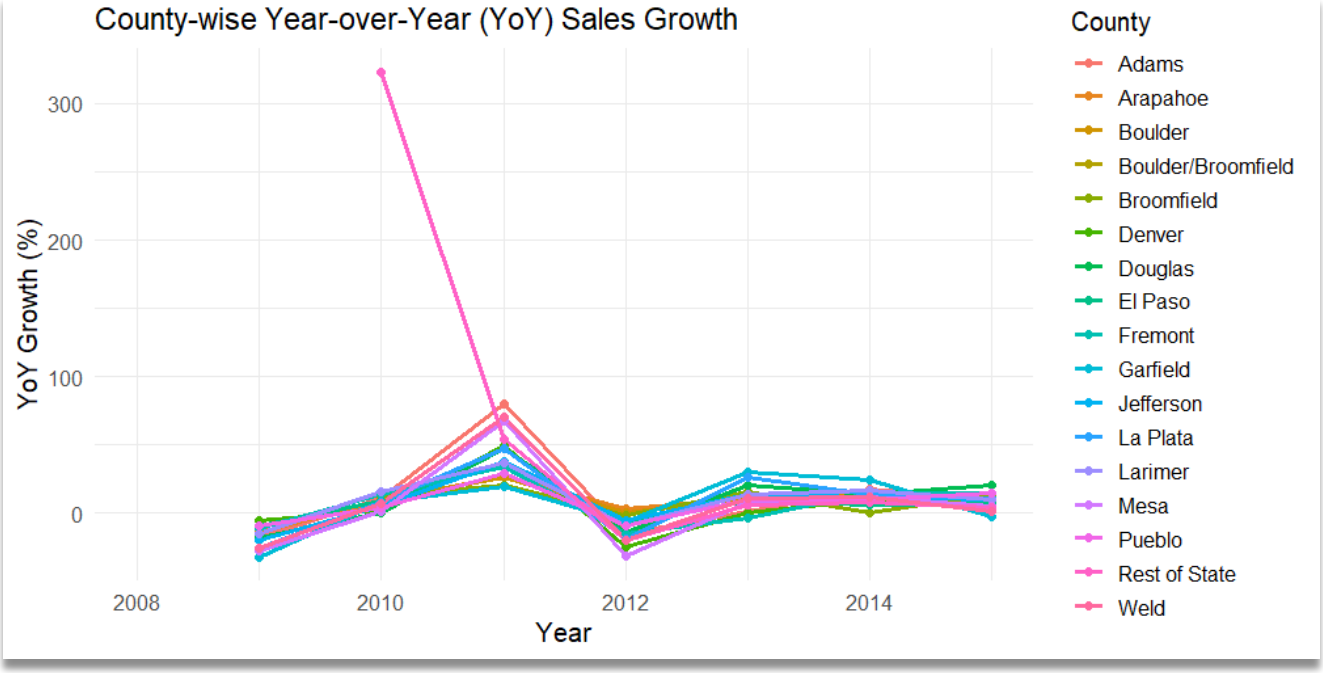
➢ Plot YoY growth(overall):

```
#Overall year over year growth in sales

yearly_sales <- df %>%
  group_by(year) %>%
  summarise(total_sales = sum(sales, na.rm = TRUE)) %>%
  arrange(year) %>%
  mutate(
    yoy_growth = (total_sales - lag(total_sales)) / lag(total_sales) * 100
  )

ggplot(yearly_sales, aes(x = year, y = yoy_growth)) +
  geom_line(color = "#0073C2FF", size = 1.2) +
  geom_point(size = 3, color = "#0073C2FF") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "gray50") +
  labs(title = "Overall Year-over-Year (YoY) Sales Growth",
       x = "Year",
       y = "YoY Growth (%)") +
  theme_minimal(base_size = 14)
```



➢ Plot YoY growth(county-wise):

```r
#County-wise YoY growth in sales

#county-wise yoy growth
county_yoy <- df %>%
  group_by(county, year) %>%
  summarise(total_sales = sum(sales, na.rm = TRUE), .groups = "drop_last") %>%
  arrange(county, year) %>%
  mutate(yoy_growth = (total_sales - lag(total_sales)) / lag(total_sales) * 100)

#plot

ggplot(county_yoy, aes(x = year, y = yoy_growth, color = county)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  labs(title = "County-wise Year-over-Year (YoY) Sales Growth",
       x = "Year",
       y = "YoY Growth (%)",
       color = "County") +
  theme_minimal(base_size = 13) +
  theme(legend.position = "right")
```



- Top and low performing counties:

```r
#Top- vs Low-Growth Counties

growth <- df %>%
  group_by(county, year) %>%
  summarise(total_sales = sum(sales)) %>%
  group_by(county) %>%
  mutate(yoy_growth = (total_sales - lag(total_sales)) / lag(total_sales))

growth_summary <- growth %>%
  summarise(avg_growth = mean(yoy_growth, na.rm = TRUE)) %>%
  arrange(desc(avg_growth))
head(growth_summary, 5)    # Top 5 counties
tail(growth_summary, 5)    # Slowest 5 counties
```

	county	avg_growth
	<chr>	<dbl>
1	Rest of State	0.625
2	Adams	0.122
3	Boulder	0.115
4	Broomfield	0.100
5	Douglas	0.0993

Top 5

	county	avg_growth
	<chr>	<dbl>
1	Rest of State	0.625
2	Adams	0.122
3	Boulder	0.115
4	Broomfield	0.100
5	Douglas	0.0993

Bottom 5

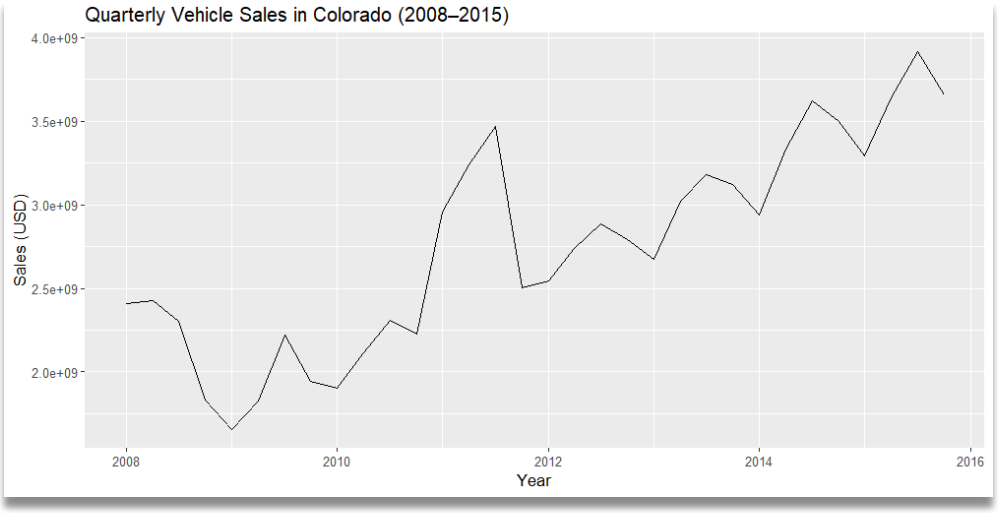
- Forecasting

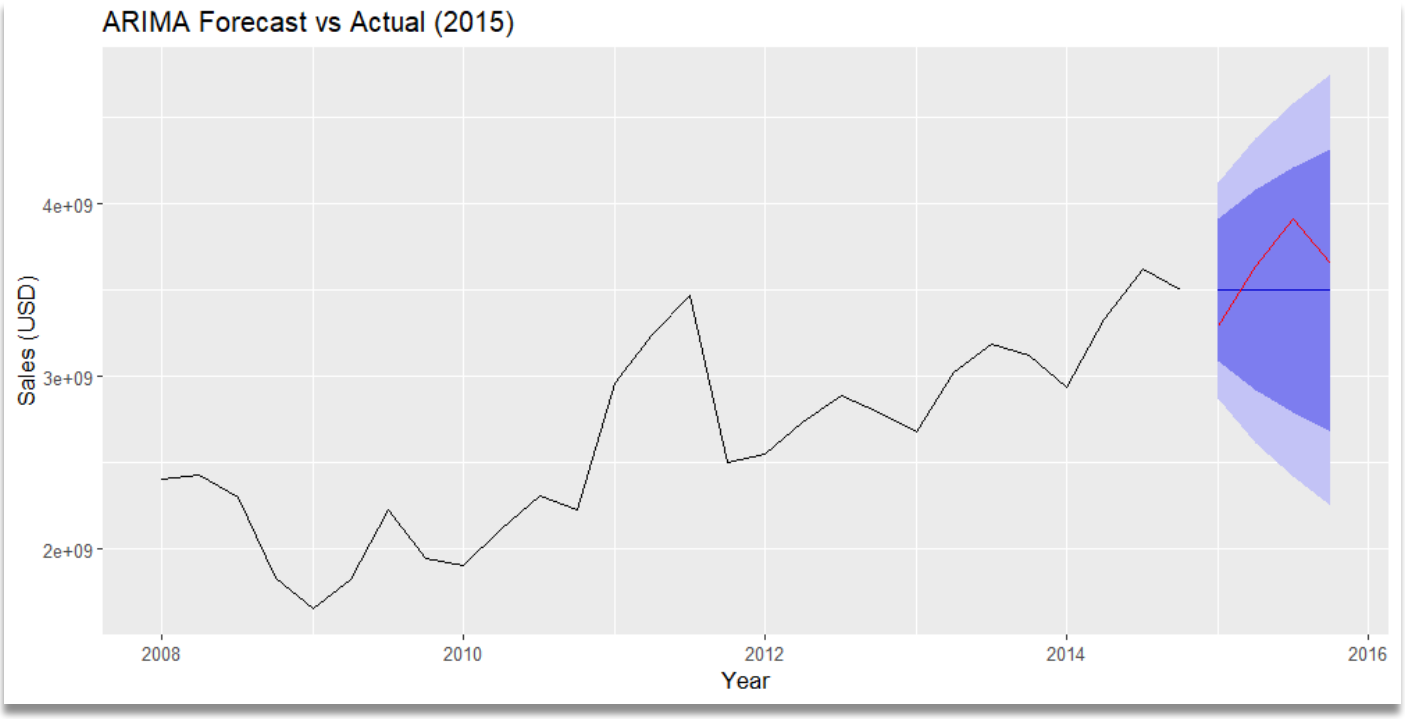
```r
#FORECAST


library(tidyverse)
library(forecast)
library(lubridate)

# Aggregate to total sales per quarter
ts_data <- df %>%
  arrange(year, quarter) %>%
  group_by(year, quarter) %>%
  summarise(total_sales = sum(sales))

# Convert to time series (quarterly)
sales_ts <- ts(ts_data$total_sales, start = c(min(ts_data$year), min(ts_data$quarter)), frequency = 4)

# Plot the time series
autoplot(sales_ts) +
  labs(title = "Quarterly Vehicle Sales in Colorado (2008-2015)",
       y = "Sales (USD)", x = "Year")

#train test

train <- window(sales_ts, end = c(2014, 4))
test  <- window(sales_ts, start = c(2015, 1))

#fit arima

fit_arima <- auto.arima(train)
summary(fit_arima)

# Forecast next 4 quarters
forecast_arima <- forecast(fit_arima, h = 4)

# Plot forecast vs actual
autoplot(forecast_arima) +
  autolayer(test, series="Actual", color="red") +
  labs(title="ARIMA Forecast vs Actual (2015)",
       y="Sales (USD)", x="Year")

accuracy(forecast_arima, test)
```

```
> summary(fit_arima)
Series: train
ARIMA(0,1,0)

sigma^2 = 1.019e+17:  log likelihood = -567.01
AIC=1136.01   AICc=1136.17   BIC=1137.31

Training set error measures:
                    ME      RMSE       MAE        MPE     MAPE      MASE        ACF1
Training set 39118234 313428413 233556805 0.5987568 9.177267 0.5451708 0.01807534
```

➤

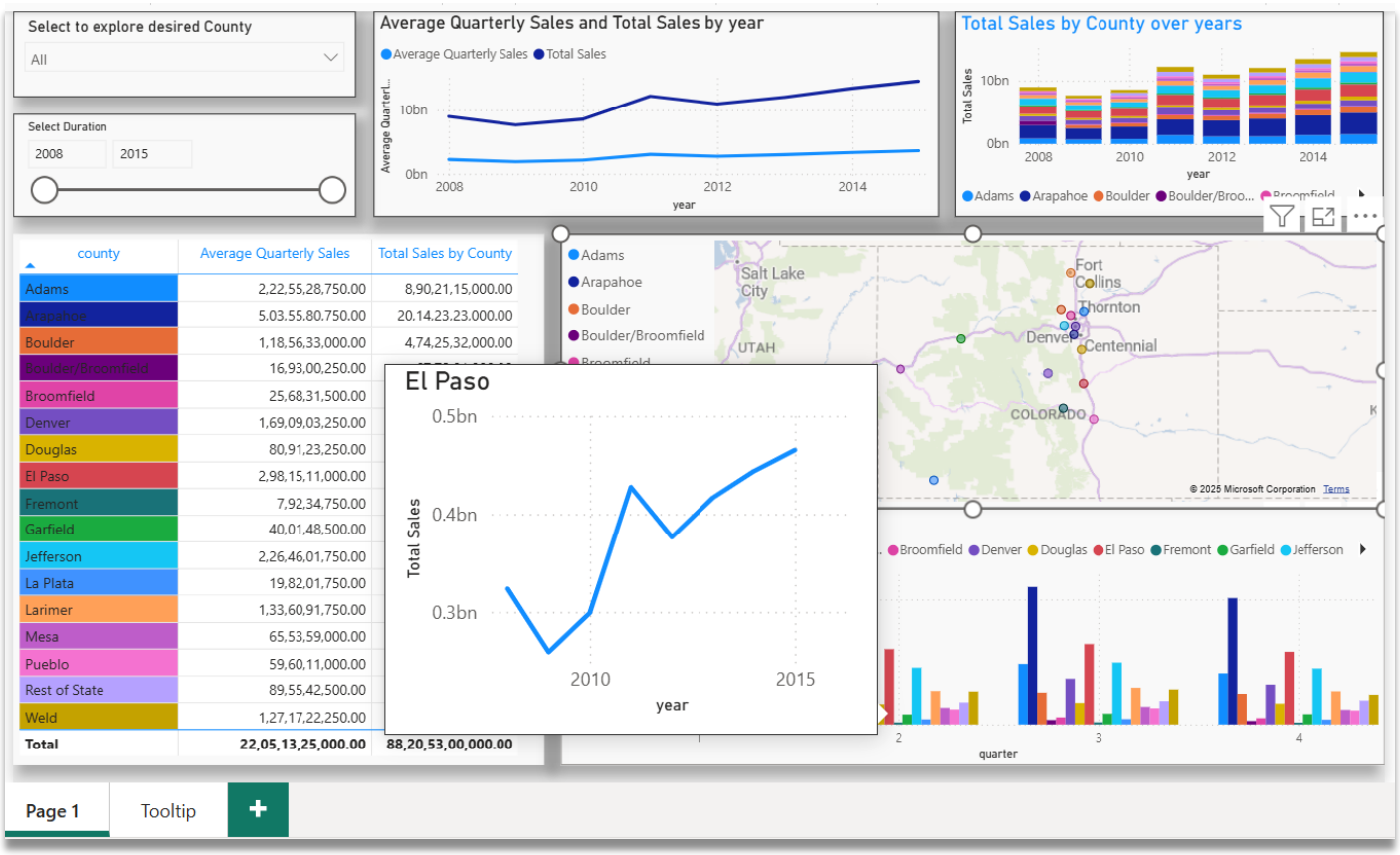
```
> accuracy(forecast_arima, test)
                    ME      RMSE       MAE        MPE     MAPE      MASE        ACF1
Training set  39118234 313428413 233556805 0.5987568 9.177267 0.5451708 0.01807534
Test set     126228000 255599414 230819000 3.1084194 6.286248 0.5387802 0.03953166
             Theil's U
Training set        NA
Test set     0.8765155
```





Chapter 5

Power BI Dashboard



Visuals used in the dashboard.

- **Line Chart to visualise Yearly Sales Trend**
- Shows and compares how average total motor vehicle sales changed from 2008 to 2015.
- Highlights major declines (e.g., recession years) and recovery periods.

- **Clustered Bar Chart to visualise County-wise Sales Comparison**
- Compares total sales across all counties over the selected period.
- Makes it easy to identify top-performing counties (e.g., Arapahoe, El Paso).

- **Clustered Column Chart to visualise County-wise Sales Comparison**
- Compares total sales across all counties over the four quarters.

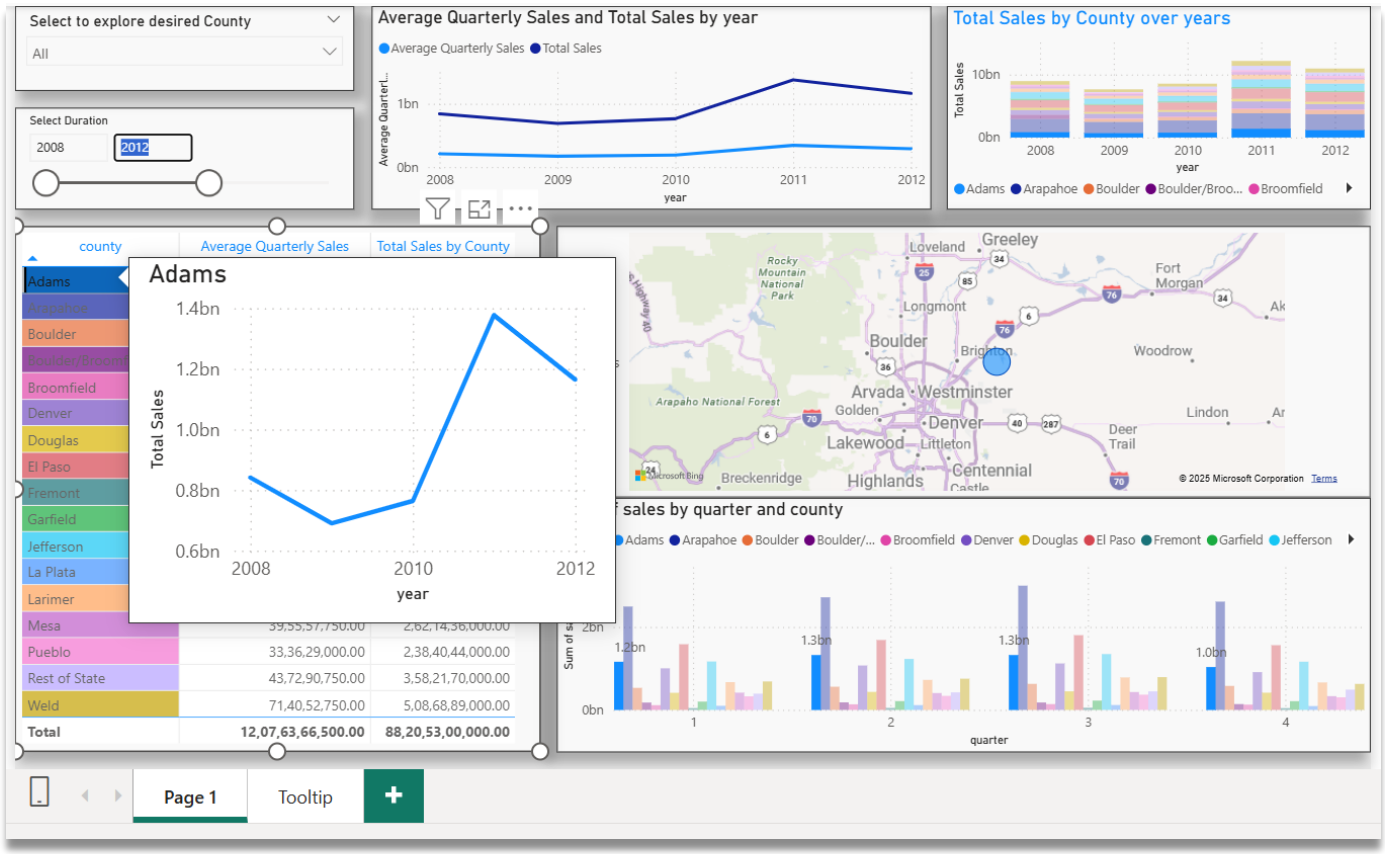
- Makes it easy to identify top-performing counties (e.g., Arapahoe, El Paso).

- **Geographic Sales Map (Filled Map / Bubble Map)**
- Visualizes county-wise sales across the state of Colorado on a map.
- Larger bubbles represent higher sales volumes.
- Helps identify geographical hotspots of vehicle sales.

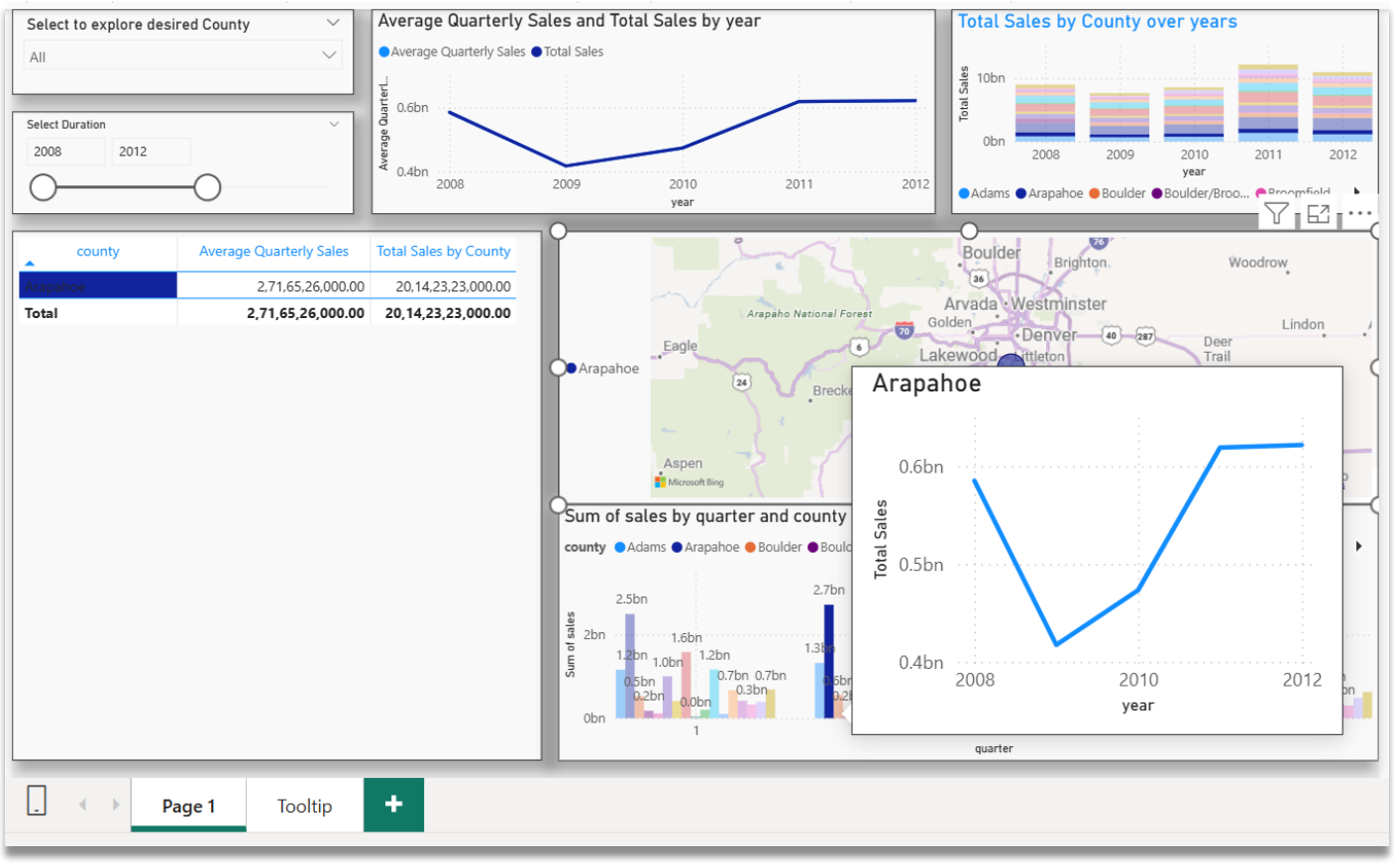
- **Detailed Sales Table (Matrix/Table Visual)**
- Lists numerical values for year, county, and total sales and average quarterly sales.
- Useful for exact value comparison and verifying aggregated trends.
- Supports deeper analysis when precise numbers are required.
- Complements visual charts by providing accurate, drill-down data.

- **Slicers (Year & County Filters)**
- Allow users to filter the dashboard by specific year(s) or county(ies).
- Makes the dashboard interactive and user-driven.
- Helps focus analysis on a particular region or time period.
- Enhances usability by supporting dynamic insights.

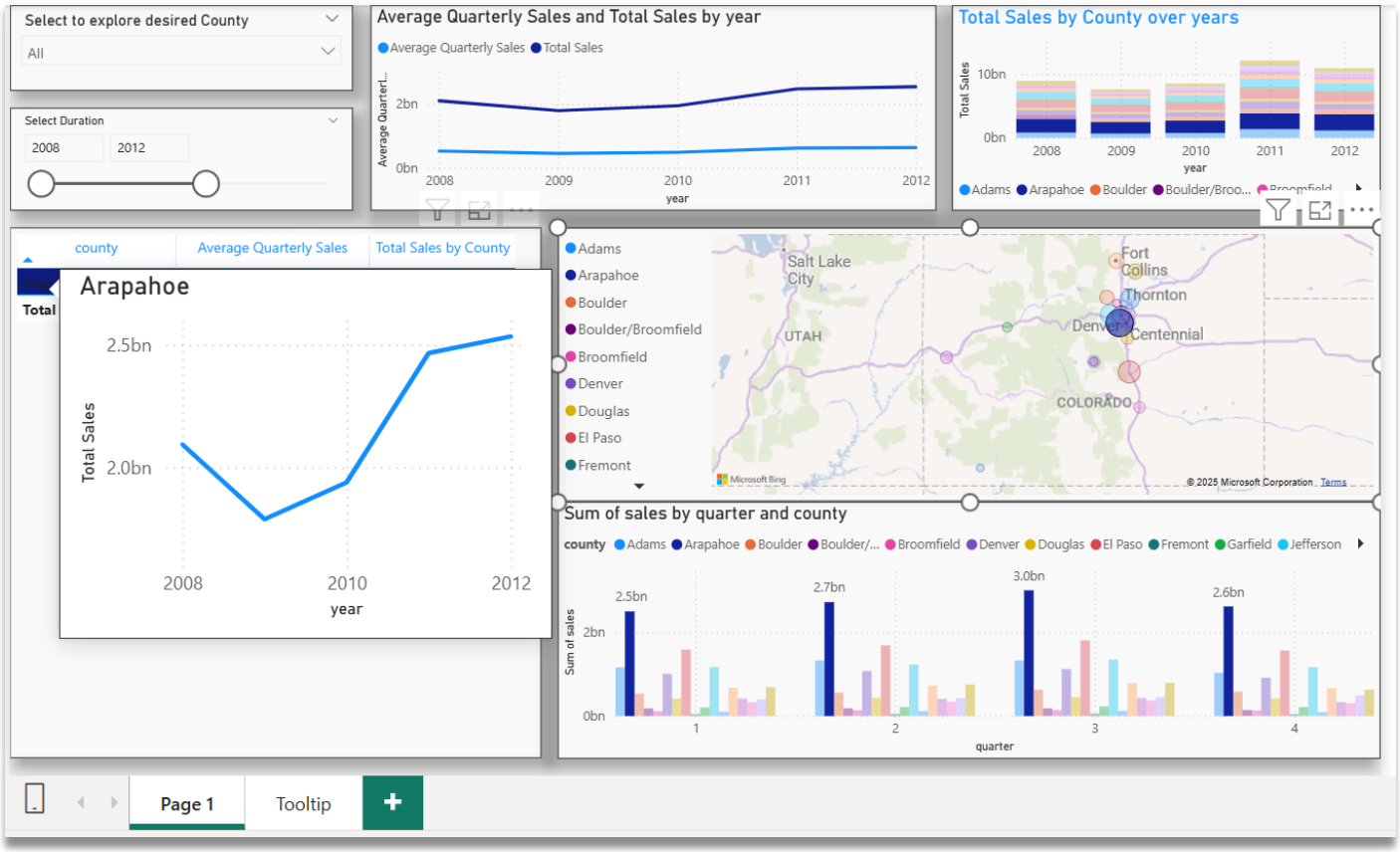
A few more screenshots from dashboard to show interactivity.



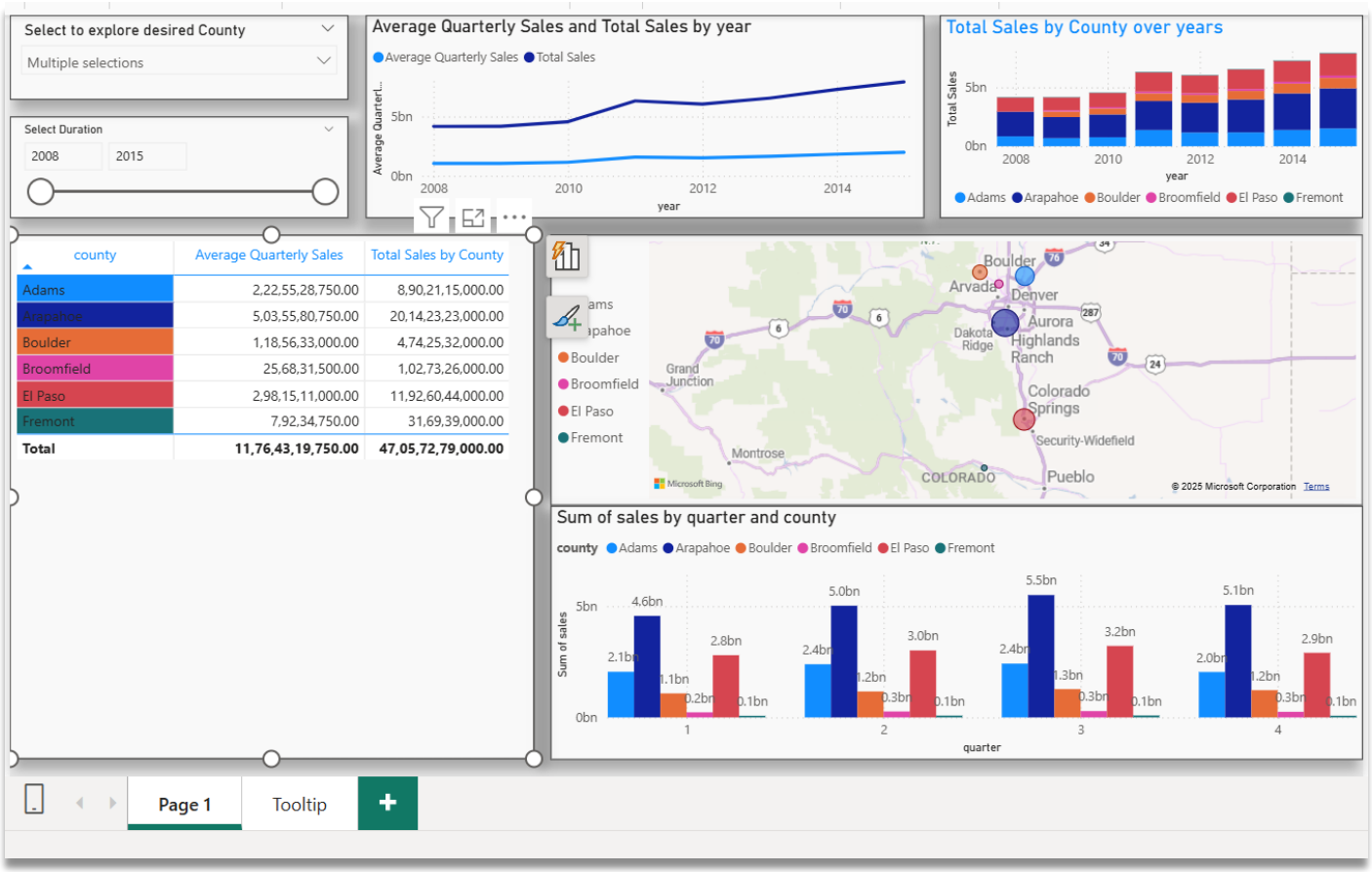
Screenshot 1 : Selected County in the year 2008-2012



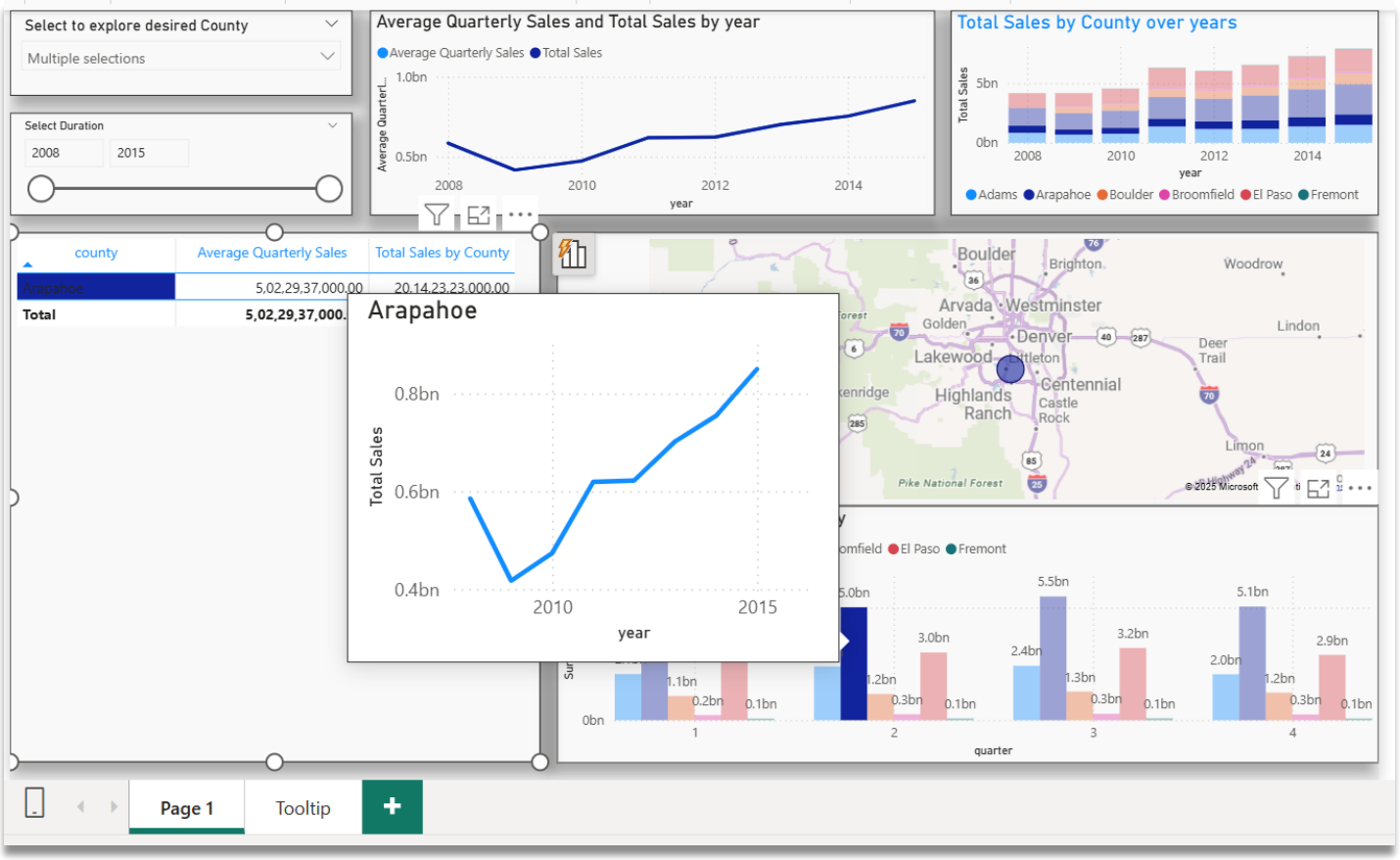
Screenshot 2: Selected County Arapahoe in 2nd quarter during the year 2008-2012



Screenshot 3: Selected County Arapahoe in all 4 quarters during the year 2008-2012



Screenshot 4: Selected Counties Adams, Arapahoe, Boulder, Broomfield, El Paso, and Fremont during the year 2008-2015



Screenshot 5: Selected County Arapahoe among the slicer selected counties above during the year 2008-2015.

Chapter 6

Results

The analysis of Colorado motor vehicle sales from 2008 to 2015 reveals several important findings. First, the statewide sales trend shows a sharp decline during the 2008–2009 recession, followed by a gradual recovery and consistent upward growth from 2011 onward. Quarterly patterns confirm clear seasonality, with **Q3 and Q4 consistently outperforming Q1 and Q2**, indicating stronger consumer demand in the second half of the year.

County-level analysis highlights strong regional disparities. **Arapahoe, El Paso, and Denver counties** emerge as the highest contributors to total statewide sales, driven by larger populations and higher economic activity. Conversely, smaller counties demonstrate limited but stable performance. The heatmap and comparative visuals further show that top-performing counties maintained their lead across most years, while lower-performing regions exhibited minimal growth.

Using ARIMA-based time series forecasting, the study predicts a continued upward trend in upcoming quarters, suggesting that Colorado's vehicle market is likely to remain strong if economic conditions remain stable. Power BI dashboard exploration validates these results, providing interactive confirmation of growth patterns, seasonal effects, and geographic variations.

Overall, the results indicate that Colorado's motor vehicle sales are influenced by both macroeconomic conditions and region-specific factors, and the forecasting model supports positive future performance.

Insights and conclusion from Dashboard

The Power BI dashboard provides a comprehensive and interactive view of Colorado's motor vehicle sales across multiple dimensions. The yearly trend analysis shows a clear decline during the 2008–2009 recession, followed by a strong and consistent recovery after 2011. County-wise comparisons reveal that urban and densely populated areas such as **Arapahoe, El Paso, and Denver** consistently outperform other regions, contributing the highest share of statewide sales. Seasonal patterns are also evident, with **Q3 and Q4 emerging as peak sales periods**, highlighting increased demand during late summer and year-end cycles. The geographic map further strengthens regional understanding by visually identifying high-performing clusters across the state. Together, these visuals confirm that Colorado's vehicle market is influenced by both macroeconomic trends and localized demographic factors. Overall, the dashboard effectively transforms raw data into actionable insights, supporting

informed decision-making related to inventory planning, regional strategy, and future sales outlook.

Conclusion

This project successfully integrates statistical analysis, time-series forecasting, and business intelligence techniques to provide a comprehensive understanding of Colorado's motor vehicle sales landscape. Through exploratory data analysis in R, key patterns such as recession-driven declines, post-2010 recovery, seasonal fluctuations, and county-wise performance differences were clearly identified. Forecasting using the ARIMA model further strengthened the analytical outcome by projecting steady growth in future sales.

The Power BI dashboard enhanced the analysis by transforming complex datasets into clear, interactive, and visually intuitive insights. It enabled users to explore year-wise, quarter-wise, and county-wise trends dynamically, making the findings accessible and business-relevant.

Overall, the study demonstrates that Colorado's motor vehicle market exhibits predictable seasonal cycles, stable long-term growth, and strong contributions from a few high-performing counties. These findings can support decision-making related to inventory planning, marketing strategy, and resource allocation for dealerships and policymakers. The combination of R-based statistical rigor and Power BI-driven visualization ensures a balanced and insightful analysis, fulfilling the objectives of the study.

Project Summary

This project presents a comprehensive analysis of Colorado's motor vehicle sales across eight years, combining R-based statistical exploration with interactive Power BI dashboards. The study begins by cleaning and structuring the dataset, followed by detailed exploratory analysis to understand year-wise, quarter-wise, and county-wise performance patterns. Clear trends emerged, including the impact of the 2008–2009 recession, subsequent recovery phases, and consistent seasonal peaks in Q3 and Q4. County-level comparisons highlighted the dominance of Arapahoe, El Paso, and Denver counties, which consistently accounted for the highest share of overall sales.

Advanced time-series forecasting using the ARIMA model provided insights into future sales performance, projecting a stable upward trajectory if economic conditions remain favourable. The Power BI dashboard served as an interactive visualization layer, enabling dynamic exploration of geographic and temporal trends. Together, these components create a unified analytical framework that transforms raw data into meaningful and actionable insights, fulfilling the objectives of the study.

Recommendations

Based on the insights derived from the data analysis, several recommendations can be made for businesses, policymakers, and stakeholders in the automotive sector:

- **Strategic Inventory Planning:**
  Given the consistently higher sales in Q3 and Q4, dealerships should ramp up inventory, marketing, and promotional campaigns ahead of these quarters.
- **Focus on High-Performing Regions:**
  Arapahoe, El Paso, and Denver counties should be prioritized for dealership expansion, enhanced service offerings, and targeted marketing due to their strong and stable sales performance.
- **Opportunity in Mid- and Low-Tier Counties:**
  Counties with moderate or lower sales can be targeted with localized campaigns, financial schemes, and mobile dealership events to stimulate demand.
- **Use Forecasting for Better Planning:**
  Forecast results predict steady growth over upcoming periods. Businesses should integrate these projections into budgeting, resource allocation, and workforce planning.
- **Enhance Customer Engagement:**
  Seasonal peaks suggest opportunities for festive promotions, loyalty rewards, and trade-in offers to maximize conversions.

Future Scope & Enhancements

- **Integration of External Variables:**
  Incorporating socio-economic factors such as population growth, fuel prices, interest rates, or income levels could provide a more holistic understanding of sales drivers.
- **More Granular Data:**
  Access to monthly or weekly sales data would allow for more precise identification of micro-seasonal trends and short-term fluctuations.
- **Categorization of Vehicle Types:**
  Including vehicle types (SUVs, sedans, EVs, trucks) can enhance market segmentation and support dealership-level decision-making.

- **Advanced Forecasting Models:**
  Future studies may explore advanced techniques such as Prophet, LSTM neural networks, or hybrid machine learning models to further improve predictive accuracy.
- **Enhanced Dashboard Functionality:**
  Adding drillthrough pages, advanced tooltips, custom bookmarks, and embedded online dashboards could make the insights more accessible and user-friendly.