Sultan Qaboos University

College of Science

Department of Statistics

Multivariate Techniques STAT5537

Fall25-Mini-Project

A Multivariate Statistical Analysis of Traffic Accidents in Oman: Exploring Gender Differences and Accident Outcomes

- **Student Name:** Manar AL Subhi
- **ID:** 137614

Contents

1. Background and Rationale

Traffic accidents are among the leading causes of injuries and fatalities worldwide, posing a major threat to public safety. Understanding the factors contributing to these accidents is crucial for developing effective prevention strategies and improving road safety policies. This study analyzes traffic accident data with a focus on gender differences and accident outcomes (injuries or deaths), as well as trends across years and among different road users such as drivers, passengers, and pedestrians. The findings aim to support evidence-based decision-making for enhancing security and safety on the roads.

2. Objectives and Research Questions

Objectives:

This study aimed to analyze and model traffic accident data to identify patterns and factors associated with the severity of accidents specifically injuries and deaths based on gender, year, and the number of drivers, passengers, and pedestrians involved. By applying some statistical methods. The study seeks to support data-driven decision-making in enhancing road safety and informing targeted prevention strategies.

Research Question:

1. Do the numbers of drivers, passengers, and pedestrians involved in accidents differ significantly across gender groups and accident outcomes (injuries vs. deaths)?
   (MANOVA)
2. Do gender groups and accident outcome groups share equal covariance structures for the road-user variables?
   (Box's M Test)
3. What is the multivariate relationship between road-user variables (drivers, passengers, pedestrians) and accident outcomes?
   (Canonical Correlation Analysis, multivariate regression)
4. Can the variability in the road-user variables be summarized effectively into a smaller number of underlying components?
   (PCA)
5. Can accident cases be accurately classified into injuries or deaths based on the road-user variables and gender?
   (Linear Discriminant Analysis)

6. How do the road-user variables change across years, and are these time patterns statistically significant?
   (MANOVA)

3. Dataset Description

This dataset derived from NCSI. Which includes data related to security and safety, most notably traffic accident cases in Oman. It contains two categorical variables—Gender (Male/Female) and Indicator (Injuries/Deaths)—as well as four numerical variables: Year, Number of Drivers, Number of Passengers, and Number of Pedestrians.

Variable	Type	Description
Year	Continuous	Year of accident
Gender	Categorical	Male/Female
Indicators	Categorical	Injuries/Deaths
Driver numbers	Continuous	Number of drivers involved
Passengers numbers	Continuous	Number of passengers involved
Pedestrian numbers	Continuous	Number of pedestrians involved

❖ **Sample Sizes per Group:**

- Male-Injuries: n = 16

- Male-Deaths: n = 12

- Female-Injuries: n = 12

- Female-Deaths: n = 12

Dataset Table:

Year	Gender	Indicators	Driver numbers	Passengers' numbers	Pedestrians numbers
2009	Male	Injuries	3811	2932	557
2009	Female	Injuries	595	1766	122
2009	Male	Deaths	407	236	157
2009	Female	Deaths	10	121	22
2010	Male	Injuries	3897	2918	595
2010	Female	Injuries	619	1938	99
2010	Male	Deaths	344	187	162
2010	Female	Deaths	15	82	30
2011	Male	Injuries	4606	3267	537
2011	Female	Injuries	767	2155	105
2011	Male	Deaths	442	299	178

2011	Female	Deaths	23	88	24
2012	Male	Injuries	4658	3307	554
2012	Female	Injuries	859	2141	99
2012	Male	Deaths	464	296	212
2012	Female	Deaths	18	117	32
2013	Male	Injuries	4499	2827	583
2013	Female	Injuries	852	1946	95
2013	Male	Deaths	381	218	187
2013	Female	Deaths	17	85	25
2014	Male	Injuries	277	535	59
2014	Female	Injuries	1616	994	354
2014	Male	Deaths	338	171	162
2014	Female	Deaths	21	102	22
2015	Male	Injuries	1692	816	299
2015	Female	Injuries	288	483	46
2015	Male	Deaths	307	156	107
2015	Female	Deaths	16	65	14
2016	Male	Injuries	1063	1293	725
2016	Female	Injuries	310	451	116
2016	Male	Deaths	280	173	142
2016	Female	Deaths	18	65	14
2017	Male	Injuries	1476	685	265
2017	Female	Injuries	248	423	37
2017	Male	Deaths	283	138	103
2017	Female	Deaths	23	79	14
2018	Male	Injuries	1280	663	244
2018	Female	Injuries	210	385	33
2018	Male	Deaths	274	150	117
2018	Female	Deaths	12	64	20
2019	Male	Injuries	1100	519	213
2019	Female	Injuries	201	370	39
2019	Male	Deaths	225	104	99
2019	Female	Deaths	9	59	15
2020	Male	Injuries	686	316	118
2020	Female	Injuries	95	129	21
2020	Male	Deaths	186	70	74
2020	Female	Deaths	8	26	7
2021	Male	Injuries	783	332	122
2021	Female	Injuries	142	216	26
2021	Male	Deaths	223	76	79
2021	Female	Deaths	13	29	14

4. Multivariate Methods

1.MANOVA (For Mean Comparison)

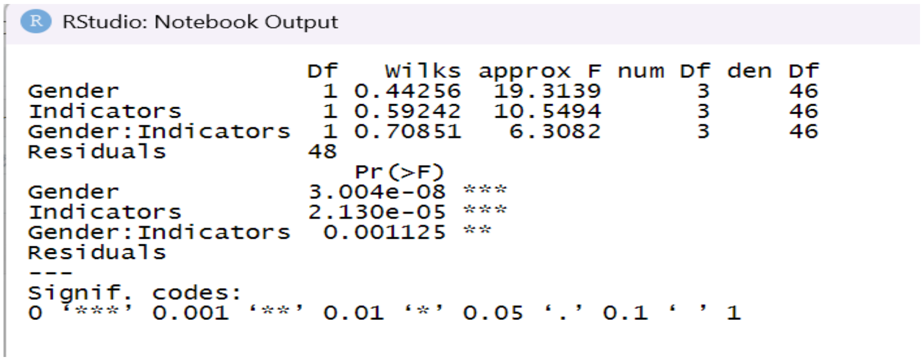
To address Research Question 1, a two-way MANOVA was conducted to test whether the multivariate mean vectors of the road-user variables (Driver numbers, Passengers numbers, Pedestrian numbers) differ across Gender and Accident Outcome (Indicators).

❖ **Formula / theoretical basis:**
Wilks' Lambda (Λ) = det(E) / det(E + H),
Where E = erros matrix, H = hypothesis matrix.

❖ **MANOVA Assumptions:**

- Multivariate normality of response variables
- Homogeneity of covariance matrices (tested via Box's M)
- Independence of observations

❖ **R-Output:**



```
                      Df   Wilks approx F num Df den Df
Gender                 1 0.44256  19.3139      3     46
Indicators             1 0.59242  10.5494      3     46
Gender:Indicators      1 0.70851   6.3082      3     46
Residuals             48
                         Pr(>F)
Gender                3.004e-08 ***
Indicators            2.130e-05 ***
Gender:Indicators      0.001125 **
Residuals
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A two-way MANOVA was conducted to examine the joint effects of Gender, Accident Outcome (Indicators: Injuries vs. Deaths), and their interaction on the multivariate response consisting of driver numbers, passengers' numbers, and pedestrian numbers.

❖ **Main Effect of Gender:**

The multivariate effect of Gender was statistically significant
(Wilks' Λ = 0.443, F (3, 46) = 19.31, $p < 0.001$).

This result indicates that the combined mean levels of drivers, passengers, and pedestrians involved in traffic accidents differ significantly between males and females.

❖ **Main Effect of Accident Outcome (Indicators):**

The multivariate effect of Accident Outcome was also statistically significant
(Wilks' Λ = 0.592, F (3, 46) = 10.55, $p < 0.001$).

This suggests that accidents resulting in injuries differ significantly from those resulting in deaths in terms of overall road-user involvement.

- **Interaction Effect: Gender × Indicators:**

The interaction between Gender and Accident Outcome was statistically significant
(Wilks' Λ = 0.709, F (3, 46) = 6.31, $p = 0.001$).

This implies that the effect of gender on road-user involvement varies depending on whether the accident outcome is an injury or a death. In other words, gender differences in accident patterns are not consistent across injury and fatal accident cases.

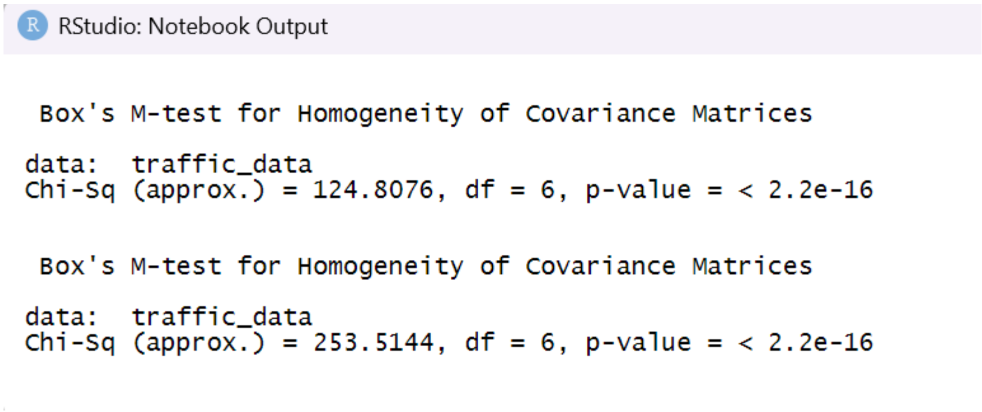
Overall Conclusion:

The MANOVA results provide strong evidence that gender, accident outcome, and their interaction jointly influence the multivariate pattern of road-user involvement. These findings justify further investigation using follow-up univariate ANOVAs or discriminant analysis to identify which specific road-user variables contribute most to the observed differences.

2. Box's M Test (For Covariance Equality)

To address Research Question 2, To assess the homogeneity of covariance matrices required for MANOVA and LDA, **Box's M Test** was applied separately for **Gender** and **Indicators**.

- **Formula/ theoretical basis:**
  Chi-square statistic to test equality of covariance matrices across groups.
- **R-Output:**



```
 Box's M-test for Homogeneity of Covariance Matrices

data:  traffic_data
Chi-Sq (approx.) = 124.8076, df = 6, p-value = < 2.2e-16

 Box's M-test for Homogeneity of Covariance Matrices

data:  traffic_data
Chi-Sq (approx.) = 253.5144, df = 6, p-value = < 2.2e-16
```

Interpretation of Box’s M Test Results

Box’s M test was conducted to assess the assumption of homogeneity of covariance matrices for the multivariate response variables (driver numbers, passengers’ numbers, pedestrian numbers) across grouping factors.

- **Box’s M Test Across Gender Groups:**

The test result was statistically significant
($\chi^2 \approx 124.81$, df = 6, $p < 0.001$).

This indicates that the covariance matrices of the road-user variables are not equal across gender groups, suggesting differences in variability and correlation structure between male and female accident cases.

- **Box’s M Test Across Accident Outcome Groups (Indicators):**

Similarly, Box’s M test for accident outcomes (Injuries vs. Deaths) was also statistically significant
($\chi^2 \approx 253.51$, df = 6, $p < 0.001$).

This result indicates a violation of the homogeneity of covariance assumption between injury and fatal accident groups.

Implications for Multivariate Analysis

The significant Box’s M test results suggest that the assumption of equal covariance matrices is violated for both Gender and Accident Outcome. However, given the relatively balanced design and moderate sample size, MANOVA using Wilks’ Lambda remains reasonably robust to such violations. Although Box’s M test indicated significant differences in covariance structures across gender and accident outcome groups, Wilks’ Lambda was used for MANOVA due to its relative robustness to violations of covariance homogeneity.

3. Canonical Correlation Analysis (CCA)

To address Research Question 3, Canonical Correlation Analysis (CCA) was initially considered to examine the multivariate relationship between the set of road-user variables (number of drivers, passengers, and pedestrians) and accident characteristics. However, Canonical Correlation Analysis requires two sets of **continuous** variables. In this study, while the road-user variables are continuous, the accident characteristics—Gender and Accident Outcome

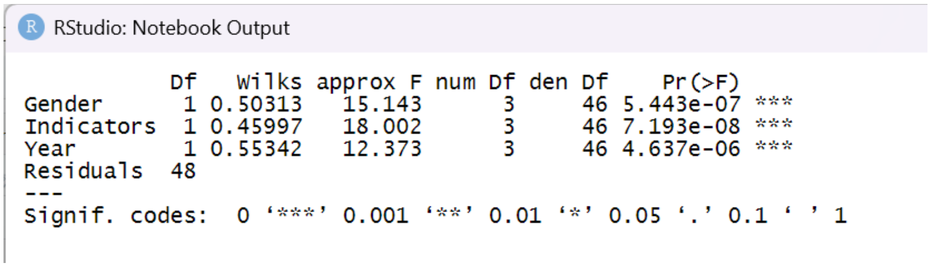
(Indicators: Injuries vs. Deaths)—are categorical. Therefore, CCA is not methodologically appropriate for inferential analysis in this context.

As a result, a **multivariate multiple regression** approach was adopted as a more suitable method to examine the joint effects of Gender, Accident Outcome, and Year on the multivariate response consisting of driver numbers, passenger numbers, and pedestrian numbers. This approach allows categorical predictors and provides valid multivariate inference using Wilks' Lambda.

Multivariate Regression

A multivariate multiple regression model was fitted with Gender, Accident Outcome (Indicators), and Year as predictors, and the three road-user variables as joint responses. Wilks' Lambda was used to assess the overall multivariate significance of each predictor.

- **Formula / theoretical basis:**
  F-statistic computed using Wilks' Lambda to assess multivariate effect of each predictor.
- **Multivariate Regression Assumptions:**
  - Multivariate normality of response variables
  - Homogeneity of covariance matrices (tested via Box's M)
  - Independence of observations
- **R-Output:**



```
           Df   Wilks approx F num Df den Df    Pr(>F)    
Gender      1 0.50313   15.143      3     46 5.443e-07 ***
Indicators  1 0.45997   18.002      3     46 7.193e-08 ***
Year        1 0.55342   12.373      3     46 4.637e-06 ***
Residuals  48                                             
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- **Effect of Gender:**

The multivariate effect of Gender was statistically significant
(Wilks' Λ = 0.503, F (3, 46) = 15.14, p < 0.001).

This result indicates that the combined numbers of drivers, passengers, and pedestrians involved in traffic accidents differ significantly between males and females, after accounting for year and accident outcome.

- **Effect of Accident Outcome (Indicators):**

The multivariate effect of Accident Outcome was also statistically significant
(Wilks' Λ = 0.460, F (3, 46) = 18.00, p < 0.001).

This suggests that injury and fatal accidents are associated with significantly different multivariate patterns of road-user involvement, even after controlling for gender and year.

❖ **<u>Effect of Year:</u>**

The multivariate effect of Year was statistically significant
(Wilks' Λ = 0.553, F (3, 46) = 12.37, $p < 0.001$).

This indicates that road-user involvement in traffic accidents has changed significantly over time, reflecting meaningful temporal trends in accident patterns across the study period.

Overall Conclusion

Overall, the multivariate regression results provide strong evidence that gender, accident outcome, and year each have a statistically significant joint effect on road-user involvement in traffic accidents. This confirms that accident characteristics in Oman are influenced by demographic factors as well as temporal changes, highlighting the importance of using multivariate methods that appropriately accommodate categorical predictors and time trends.

4. Principal Component Analysis (PCA)

To address Research Question 4, to reduce dimensionality and identify underlying patterns among the road-user variables, Principal Component Analysis (PCA) was performed using standardized variables. Where Principal components explaining the majority of total variance were retained and interpreted based on their loading structures.

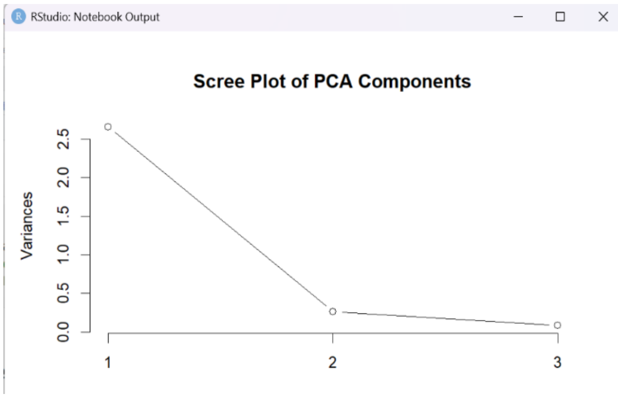
❖ **<u>Formula / theoretical basis:</u>**
Principal components are eigenvectors of the covariance/correlation matrix, with variance explained by corresponding eigenvalues.

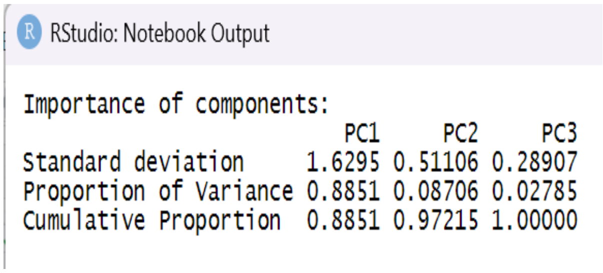
❖ **<u>PCA Assumptions:</u>**

- Variables are linearly related
- Large enough sample size for stable component estimation
- Continuous variables

PCA assumptions were examined. The road-user variables showed moderate-to-strong linear correlations, justifying dimensionality reduction. The Kaiser-Meyer-Olkin (KMO) measure indicated sampling adequacy, and Bartlett's test of sphericity was significant, supporting the suitability of the dataset for PCA.

❖ **<u>R-Output:</u>**





```
Importance of components:
                          PC1     PC2     PC3
Standard deviation     1.6295 0.51106 0.28907
Proportion of Variance 0.8851 0.08706 0.02785
Cumulative Proportion  0.8851 0.97215 1.00000
```

- **PC1** has a standard deviation of 1.63 and explains **88.5% of the total variance**. And it captures the dominant pattern of variation across drivers, passengers, and pedestrians involved in accidents. Indicating that driver, passenger, and pedestrian numbers tend to increase or decrease together.
- **PC2** explains an additional **8.7% of the total variance**.
- **PC3** explains **2.8% of the total variance**.

PC2 and **PC3** contribute very little additional information, suggesting that the dataset is highly correlated and can be effectively summarized with the first component or the first two components.

The first principal component explains 88.5% of the variance in road-user involvement, indicating that the three variables are highly correlated and can be effectively summarized using one or two components where the first two principal components together account for 97.2% of the total variance, indicating that most of the variability in road-user involvement can be summarized with just two components.

❖ **Scree Plot Interpretation**
The scree plot of the principal components visually confirms the PCA results. The first component (PC1) shows a much higher variance compared to PC2 and PC3, indicating that it captures the dominant pattern of variation in the road-user variables (driver, passenger, and pedestrian numbers). PC2 and PC3 contribute only minor additional variance, consistent with their explained proportions of 8.7% and 2.8%, respectively. This sharp drop in variances after PC1 confirms that most of the variability in road-user involvement can be effectively summarized using just the first component, or the first two components if near-complete variance retention is desired. The scree plot thus supports the dimensionality reduction interpretation and reinforces the high correlation among the three road-user variables.

This dimensionality reduction simplifies subsequent analysis, visualization, and interpretation, while retaining nearly all the variability in the original variables.

5. Linear Discriminant Analysis (LDA)

To address Research Question 5, Linear Discriminant Analysis was used to classify accident cases into Injuries or Deaths based on road-user variables (driver numbers, passengers' numbers, pedestrian numbers) and Gender.

❖ **Formula / theoretical basis:**

Discriminant function: Y = w1*X1 + w2*X2 + ... + wp*Xp,

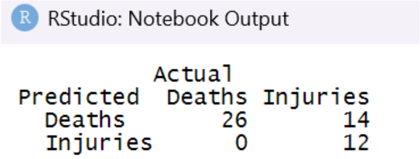
where w are weights maximizing separation between groups.

❖ **LDA Assumptions:**

- Multivariate normality within groups
- Equality of covariance matrices between groups (moderately violated but method is robust)

❖ **R-Output:**

Confusion Matrix:



RStudio: Notebook Output

```
          Actual
Predicted  Deaths Injuries
  Deaths       26       14
  Injuries      0       12
```

- **Correctly classified Deaths:** 26 out of 40 → 65%
- **Correctly classified Injuries:** 12 out of 12 → 100%
- Overall, the model shows good classification performance, especially for **Injuries**, but some Death cases were misclassified as Injuries (14 cases), indicating that road-user variables and gender alone do not fully explain fatal accident occurrence. This suggests that additional predictors or more complex models may improve classification performance.

Classification Accuracy:

$$\text{Overall Accuracy} = \frac{26 + 12}{26 + 14 + 0 + 12} = \frac{38}{52} \approx 73\%$$

LDA can effectively predict accident outcomes using road-user variables and gender. Misclassification of some Death cases suggests that additional predictors or more complex models may improve classification performance. The LDA model correctly classified 73% of traffic accidents, performing perfectly for Injuries but misclassifying some Death cases, indicating moderate predictive accuracy.

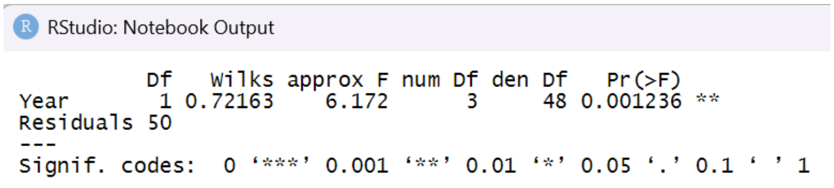
6. Multivariate Analysis of Time Trends (MANOVA)

To address Research Question 6, A multivariate analysis of variance (MANOVA) was conducted to examine whether the multivariate pattern of road-user involvement (driver numbers, passenger numbers, and pedestrian numbers) differed significantly across years.

❖ **Assumptions:**

- Multivariate normality of response variables
- Homogeneity of covariance matrices (tested via Box's M)
- Observations are independent across years

❖ **R-Output:**

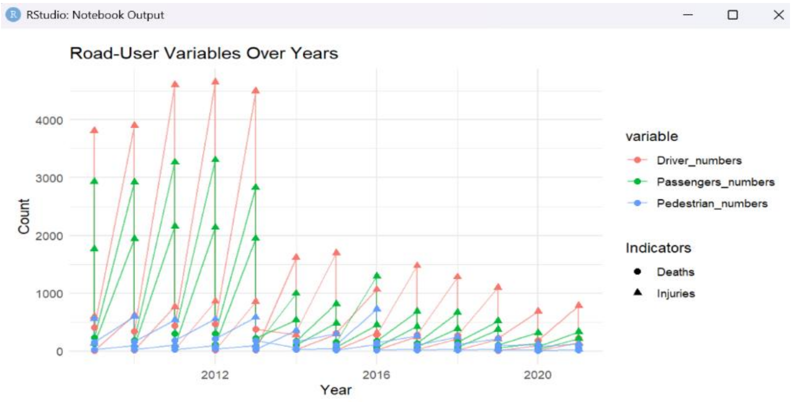


```
          Df   Wilks approx F num Df den Df   Pr(>F)   
Year       1 0.72163    6.172      3     48 0.001236 **
Residuals 50                                           
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Effect of Year
The multivariate effect of Year was statistically significant
(Wilks' Λ = 0.722, F(3, 48) = 6.17, p = 0.001).

This result indicates that the combined numbers of drivers, passengers, and pedestrians involved in traffic accidents changed significantly over time. In other words, there are meaningful multivariate time trends in road-user involvement across the study period.

The significant effect of year suggests that traffic accident patterns in Oman are not stable over time, likely reflecting changes in traffic volume, road safety policies, enforcement measures, or population behavior. These temporal variations highlight the importance of incorporating time effects when analyzing traffic accident data.



As illustrated in the figure, the number of drivers (red line) consistently exceeds that of passengers (green line) and pedestrians (blue line), with noticeable peaks around 2012–2013 followed by a general decline in later years. Both deaths (dots) and injuries (triangles) exhibit similar temporal patterns, indicating that accident outcomes vary alongside road-user involvement. The visual evidence strongly supports the MANOVA results and confirms the presence of significant multivariate temporal trends in traffic accident characteristics.

5. Overall Conclusion

This study applied a suite of multivariate statistical methods to analyze traffic accident data in Oman, focusing on gender differences, accident outcomes, and temporal trends in road-user involvement. The key findings are as follows:

1. **Gender and Accident Outcomes**:
   MANOVA results revealed that the combined numbers of drivers, passengers, and pedestrians differ significantly across gender and accident outcome groups. Furthermore, the interaction between gender and accident outcome was significant, indicating that gender differences in accident patterns vary depending on whether the outcome is an injury or a death.

2. **Covariance Structures**:
   Box's M Test indicated that the covariance matrices of road-user variables differ significantly across gender and accident outcome groups. While this violates the homogeneity assumption, the MANOVA results using Wilks' Lambda remain robust, providing reliable inference.

3. **Multivariate Relationships**:
   Although Canonical Correlation Analysis was initially considered, it was not appropriate due to categorical accident outcome variables. Multivariate multiple regression revealed that Gender, Accident Outcome, and Year each have a significant joint effect on driver, passenger, and pedestrian numbers, confirming meaningful associations between demographic, outcome, and temporal factors.

4. **Dimensionality Reduction**:
   PCA indicated that the first two principal components capture over 97% of the variability in road-user involvement, suggesting that the three variables are highly correlated. This allows for effective simplification in further analyses and visualizations.

5. **Classification of Accident Outcomes**:
   LDA showed that accident cases can be classified into Injuries or Deaths with an overall accuracy of 73%. The model performed perfectly for Injuries but misclassified some Death cases, indicating moderate predictive power and highlighting the potential benefit of including additional predictors.

6. **Time Trends**:
   Multivariate analysis of time trends (MANOVA) demonstrated that road-user involvement has changed significantly over the years. This reflects temporal variations likely influenced by changes in traffic volume, road safety policies, enforcement measures, and population behavior.

❖ **Limitations**:

- Canonical Correlation Analysis was not suitable due to categorical outcome variables.
- Box's M test indicated covariance inequality; MANOVA is robust but results should be interpreted with caution.
- LDA misclassified some Death cases, suggesting that other variables (road type, vehicle type, environmental factors) may improve prediction.
- Sample sizes for some groups are small, limiting generalizability.

❖ **Recommendations / Implications:**

- Include additional predictors (road type, weather conditions, vehicle type) in future analyses.
- Develop gender-specific and injury/fatality-specific road safety interventions.
- Consider policies targeting high-risk years or periods based on observed time trends.

Overall, these analyses confirm that traffic accident patterns in Oman are influenced by gender, accident outcomes, and temporal factors. Road-user involvement is highly interrelated, allowing dimensionality reduction without significant loss of information. Classification and multivariate analyses provide actionable insights for identifying high-risk groups and developing targeted safety interventions. The findings emphasize the importance of applying appropriate multivariate methods that accommodate both continuous and categorical predictors to accurately capture complex patterns in traffic accident data.