B9DA101 Statistics for Data Analytics: CA_ONE

January 2025

Submitted by:

Sharath Chandra Chavali: 20058721

Anish Rao: 20066423

Adithya Durgapu: 20054683

Shruthi Ravi: 20068068

Lecturer: Dr. Muhammad Alli

Index

Introduction

Purpose of the Analysis

This analysis considers a real-world relational dataset that contains at least two categorical and two continuous variables. The objective is to explore the dataset using descriptive analytics, apply probability models to quantify uncertainty, and perform hypothesis testing to validate statistical assumptions.

Dataset Overview

For this analysis, the "**Diamonds"** dataset (a built-in dataset in R) is used. This dataset consists of 53,940 observations, with attributes describing diamond quality, size, and price. It includes both categorical and continuous variables, making it ideal for statistical exploration.

Key Variables in the Dataset:

- Continuous Variables:
  - price: Price of the diamond in US dollars.
  - carat: Weight of the diamond in carats.
- Categorical Variables:
  - cut: Quality of the cut (Fair, Good, Very Good, Premium, Ideal).
  - color: Diamond color, graded from D (best) to J (worst).

Structure of Dataset

```
> str(diamonds)
tibble [53,940 × 10] (S3: tbl_df/tbl/data.frame)
 $ carat  : num [1:53940] 0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
 $ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...
 $ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...
 $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...
 $ depth  : num [1:53940] 61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
 $ table  : num [1:53940] 55 61 65 58 58 57 57 55 61 61 ...
 $ price  : int [1:53940] 326 326 327 334 335 336 336 337 337 338 ...
 $ x      : num [1:53940] 3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
 $ y      : num [1:53940] 3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
 $ z      : num [1:53940] 2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

Google Colab Notebook link-
https://colab.research.google.com/drive/1erHJ0YDf0VXzdQWUhKeesk_CtijXh46L?usp=sharing

Q1. Descriptive Analytics & Outlier Detection

(a) Data Visualization & Description

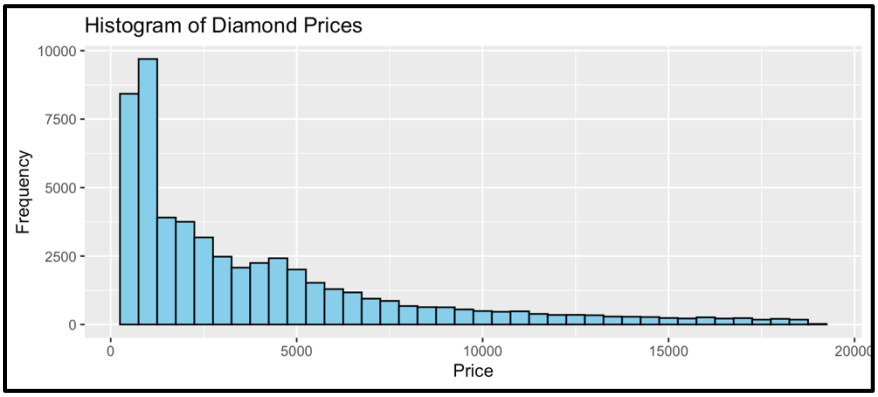
Describe the dataset using appropriate plots/curves/charts, ... (7)

Code -

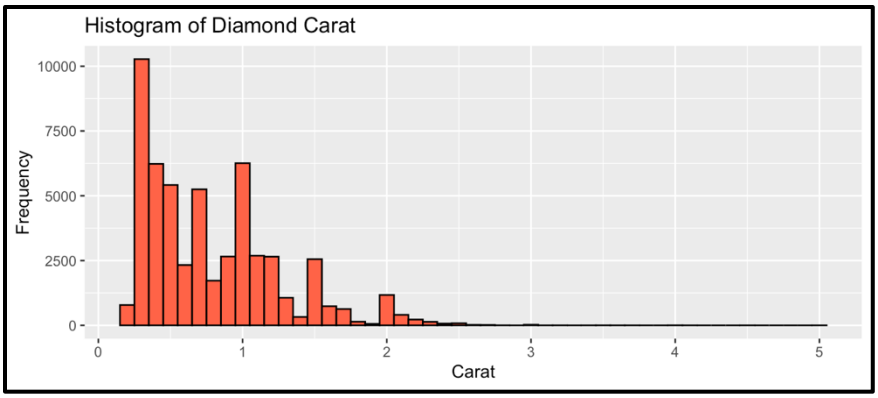
```
ggplot(diamonds, aes(x = price)) +
  geom_histogram(binwidth = 500, fill = "skyblue", color = "black") +
  labs(title = "Histogram of Diamond Prices", x = "Price", y = "Frequency")

ggplot(diamonds, aes(x = carat)) +
  geom_histogram(binwidth = 0.1, fill = "tomato", color = "black") +
  labs(title = "Histogram of Diamond Carat", x = "Carat", y = "Frequency")

ggplot(diamonds, aes(x = cut)) +
  geom_bar(fill = "lightgreen") +
  labs(title = "Distribution of Diamond Cut", x = "Cut", y = "Count")

ggplot(diamonds, aes(x = color)) +
  geom_bar(fill = "pink") +
  labs(title = "Distribution of Diamond Color", x = "Color", y = "Count")
```

The code generates histograms for continuous variables (price, carat) and bar charts for categorical variables (cut, color) to understand their distributions.

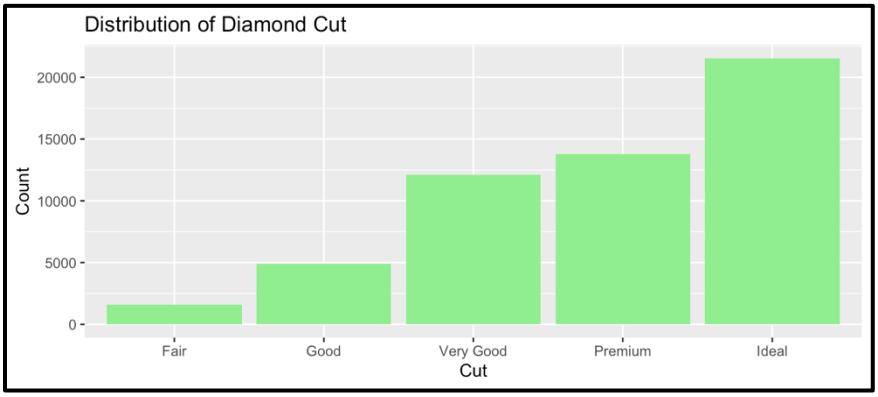
Output -



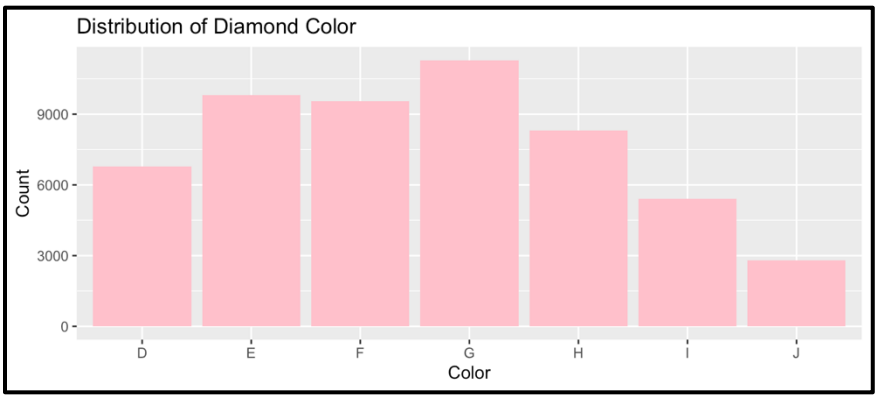
Price distribution is right-skewed, meaning most diamonds are lower-priced, with a few high-price ones.



Carat distribution is also right-skewed, indicating most diamonds are small.



Cut Distribution: Majority of diamonds have an Ideal cut.



Color Distribution: Some color grades (E, F, G) are more common than others.

(b) Central & Variational Measures

Consider one of continuous attributes, and compute central and variational measures (8)

Code -

```
price = diamonds$price
price_mean = mean(price)
price_median = median(price)
price_range = range(price)
price_variance = var(price)
price_sd = sd(price)
price_IQR = IQR(price)

cat("Central Measures for Price: \nMean:", price_mean,
    "\nMedian:", price_median,
    "\n\nVariational Measures for Price: \nRange:",
    price_range[1], "to", price_range[2],
    "\nVariance:", price_variance,
    "\nStandard Deviation:", price_sd,
    "\nIQR:", price_IQR)
```

The code calculates Central measures for price - mean, median, and Variational measures - range, variance, standard deviation, and interquartile range (IQR).

Output -

```
Central Measures for Price:
Mean: 3932.8
Median: 2401

Variational Measures for Price:
Range: 326 to 18823
Variance: 15915629
Standard Deviation: 3989.44
IQR: 4374.25
```

Insights-

- Mean Price = 3932.80, Median Price = 2401.00. Since mean is greater than median, our dataset has a Right-skewed distribution.
- Price Range between 326 and 18823 shows a widespread in diamond prices.
- Standard Deviation of 3989.44 shows a high variability in price.

(c) Chebyshev's Rule & Outliers

For a particular variable of the dataset, use Chebyshev's rule, and propose one-sigma interval. Based on your proposed interval, specify the outliers if any. (10)

Chebyshev's theorem states that for any probability distribution, at least $1 - \frac{1}{k^2}$ of the data falls within k standard deviations $(k\sigma)$ from the mean (μ) for k >1.

- For k = 2 (two standard deviations), at least 75% of the data is expected to be within $\mu \pm 2\sigma$.

Code -

```
lower_bound = price_mean - price_sd
upper_bound = price_mean + price_sd
cat("One-sigma interval (Mean ± SD): [", lower_bound, ",", upper_bound, "]\n")

price_outliers = diamonds[diamonds$price < lower_bound
                          | diamonds$price > upper_bound, ]
cat("Number of outliers:", nrow(price_outliers), "\n")
```

The code applies Chebyshev's Rule to compute a one-sigma interval (Mean ± SD) for price and counts data points falling outside this range as outliers.

Output-

```
One-sigma interval (Mean ± SD): [ -56.64002 , 7922.239 ]
Number of outliers: 7715
```

Insights-

- One-Sigma Interval: [-56.64, 7922.24]
- There are 7715 outliers all over the price of 7922.24 confirming a right-skewed distribution.

(d) Boxplot Method for Outlier Detection

Explain how the box-plot technique can be used to detect outliers. Apply this technique for one attribute of the dataset (10)

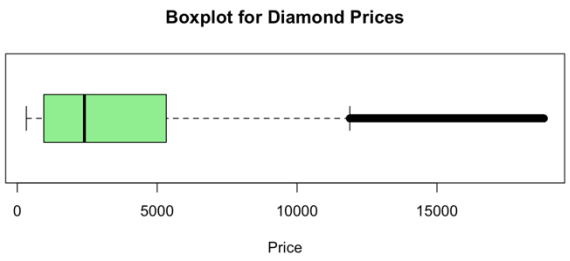
Code -

```
first_quartile = quantile(diamonds$price, 0.25)
third_quartile = quantile(diamonds$price, 0.75)
price_IQR = IQR(diamonds$price)

lower_limit = first_quartile - 1.5 * price_IQR
upper_limit = third_quartile + 1.5 * price_IQR

price_outlier_values = diamonds$price[diamonds$price < lower_limit |
                                      diamonds$price > upper_limit]
num_outliers = length(price_outlier_values)
if(num_outliers == 0) {
  cat("No outliers detected using the boxplot method.\n")
} else {
  cat("Outliers detected using the boxplot method:", num_outliers, "\n")
}

boxplot(diamonds$price,
        main = "Boxplot for Diamond Prices",
        col = "lightgreen",
        horizontal = TRUE,
        xlab = "Price")
```

The code uses the IQR method (1.5 × IQR rule) to detect outliers in price, identifying values beyond Q1 - 1.5×IQR and Q3 + 1.5×IQR. A boxplot is also generated for visualization.

Output-

```
Outliers detected using the boxplot method: 3540
```



Insights-

- Outlier Threshold (Upper Limit): Prices above 11886 are detected as outliers.
- 3540 observations exceed this threshold.
- Boxplot Interpretation:
- The black dots beyond the right whisker represent high-priced outliers.

Q2. Descriptive Analytics & Outlier Detection

(a) Proposed Probability Models

Select four variables of the dataset and propose an appropriate probability model to quantify uncertainty of each variable. (10)

To quantify uncertainty in the dataset, we select two **continuous - price, carat** and two **categorical - cut, color** variables.

Probability Model Selection:

Normal Distribution (For Continuous Variables - price, carat)

- The Normal distribution is commonly used to model continuous numerical variables that vary around an average.
- Price and carat both exhibit natural variations, making the Normal distribution a suitable choice for understanding their central tendency and dispersion.
- The Normal model helps in making statistical inferences and probability estimations for these variables.

Multinomial Distribution (For Categorical Variables - cut, color)

- The Multinomial distribution is appropriate for categorical variables with multiple distinct categories.
- Cut (Fair, Good, Very Good, Premium, Ideal) and color (D to J) both fall into this category
- This model allows us to estimate the probability of each category occurring, helping in uncertainty quantification for these attributes.

(b) Estimation of Model Parameters

For each model in part (a), estimate the parameters of model (10)

Code -

```
# Continuous Variables - Normal Model
price_mean = mean(diamonds$price)
price_sd = sd(diamonds$price)

carat_mean = mean(diamonds$carat)
carat_sd = sd(diamonds$carat)

cat(" Normal distribution parameters for Price:\n",
    "Mean:", price_mean, "\n", "SD:", price_sd, "\n\n",
    "Normal distribution parameters for Carat:\n",
    "Mean:", carat_mean, "\n", "SD:", carat_sd, "\n\n")

# Categorical Variables - Multinomial Model
cut_freq = table(diamonds$cut)
cut_prob = prop.table(cut_freq)

color_freq = table(diamonds$color)
color_prob = prop.table(color_freq)

cat("Estimated probabilities for Cut:\n"); print(cut_prob)
cat("\nEstimated probabilities for Color:\n"); print(color_prob)
```

The code calculates parameters for each probability model, estimating mean and standard deviation for continuous variables (price, carat) and category probabilities for categorical variables (cut, color).

Output-

```
 Normal distribution parameters for Price:
 Mean: 3932.8
 SD: 3989.44

 Normal distribution parameters for Carat:
 Mean: 0.7979397
 SD: 0.4740112
```

```
Estimated probabilities for Cut:

      Fair       Good  Very Good    Premium      Ideal
0.02984798 0.09095291 0.22398962 0.25567297 0.39953652


Estimated probabilities for Color:

         D          E          F          G          H          I          J
0.12560252 0.18162773 0.17690026 0.20934372 0.15394883 0.10051910 0.05205784
```

Insights-

Continuous Variables (price, carat)

- Mean Price: 3932.80, SD: 3989.44, Indicates high variation in diamond prices.
- Mean Carat: 0.80, SD: 0.47 Indicates most diamonds are small, with slight variation in weight.

Categorical Variables (cut, color)

- Ideal cut is the most common (39.95%), suggesting a preference for high-quality cuts.
- G is the most frequent color (20.93%), indicating its popularity in diamonds.
- These probabilities help estimate the likelihood of cuts and colors in future samples.

(c) Predictive Analytics & Predictions

Express the way in which each model can be used for the predictive analytics, then find the prediction for each attribute (15)

Code -

```
# Predictions for Continuous Variables
predicted_price = price_mean
predicted_carat = carat_mean
cat("Predicted Price (expected value):", predicted_price,
    "\nPredicted Carat (expected value):", predicted_carat, "\n\n")

# Predictions for Categorical Variables
predicted_cut = names(which.max(cut_prob))
predicted_color = names(which.max(color_prob))
cat("Predicted Cut (most common):", predicted_cut,
    "\nPredicted Color (most common):", predicted_color)
```

The code predicts expected values for continuous variables (price, carat) and identifies the most probable categories for categorical variables (cut, color).

Output-

```
Predicted Price (expected value): 3932.8
Predicted Carat (expected value): 0.7979397
```

```
Predicted Cut (most common): Ideal
Predicted Color (most common): G
```

Insights-

Continuous Variables (price, carat)

- Predicted Price: 3932.80 (mean price)
- Predicted Carat: 0.80 (mean weight)
- The mean is used as the best estimate for expected values in new samples.

Categorical Variables (cut, color)

- Most common cut = Ideal (39.95%). Likely to appear most frequently.
- Most common color = G (20.93%). Highest probability in the dataset.
- These predictions help estimate likely outcomes in future diamonds.

Q3. Hypothesis Testing

(a) Chi-Square Test for Independence

Consider two categorical variables of the dataset, develop a binary decision-making strategy to check whether two variables are independent at the significant level alpha=0.01. To do so, (10)

(i) State the hypotheses.

Null Hypothesis (H_o): The variables Cut and Color are independent (no relationship between them).

Alternative Hypothesis (H_A): The variables Cut and Color are dependent (there is a relationship between them).

(ii) Find the statistic and critical values.

Code-

```
contingency_table = table(diamonds$cut, diamonds$color)
chi_test = chisq.test(contingency_table)

chi_stat = chi_test$statistic
df = chi_test$parameter

# critical value = qchisq(1 - alpha, df)
critical_value = qchisq(0.99, df)

cat("Chi-square Statistic:", chi_stat,
    "\nDegrees of Freedom:", df,
    "\nCritical Value (alpha = 0.01):", critical_value, "\n")
```

```
if (chi_stat > critical_value) {
  cat("Reject H0: chi_stat =", round(chi_stat, 2),
      ">", round(critical_value, 2),
      "; Cut and Color are dependent.\n")
} else {
  cat("Fail to reject H0: chi_stat =", round(chi_stat, 2),
      "<=", round(critical_value, 2),
      "; insufficient evidence of dependency.\n")
}
```

The code creates a contingency table for cut and color, then performs a Chi-Square test. It extracts the Chi-Square statistic, degrees of freedom, and critical value

(iii) Explain your decision and Interpret results.

Output-

```
Chi-square Statistic: 310.3179
Degrees of Freedom: 24
Critical Value (alpha = 0.01): 42.97982
```

```
Reject H0: chi_stat = 310.32 > 42.98 ; Cut and Color are dependent.
```

Insights-

- Chi-Square Statistic (310.32) > Critical Value (42.98), so we reject H_O .
- Cut and Color are dependent.

(b) Goodness-of-Fit Test

Consider one categorical variable, apply goodness of fit test to evaluate whether a candidate set of probabilities can be appropriate to quantify the uncertainty of class frequency at the significant level alpha=0.05. (10)

Hypothesis -

The Goodness-of-Fit Test checks whether the observed distribution of a categorical variable (cut) follows an assumed probability distribution.

Null Hypothesis (H_O): The observed frequencies of cut follows the expected probabilities.

Alternative Hypothesis (H_A): They do not follow the expected probabilities.

Code -

```r
observed = table(diamonds$cut)
n_levels = length(observed)
candidate_prob = rep(1/n_levels, n_levels)

gof = chisq.test(observed, p = candidate_prob)
gof_stat = gof$statistic
df = gof$parameter
critical_value = qchisq(0.95, df)

cat("Chi-square Statistic:", round(gof_stat, 2),
    "\nDegrees of Freedom:", df,
    "\nCritical Value (alpha = 0.05):", round(critical_value, 2), "\n")

if (gof_stat > critical_value) {
  cat("Reject H0: gof_stat =", round(gof_stat, 2),
      ">", round(critical_value, 2),
      "; the candidate probabilities do not fit the data.\n")
} else {
  cat("Fail to reject H0: gof_stat =", round(gof_stat, 2),
      "<=", round(critical_value, 2),
      "; the candidate probabilities are appropriate.\n")
}
```

The code creates a frequency table for cut, assumes an equal probability for each category, and performs a Chi-Square Goodness-of-Fit test. It extracts the Chi-Square statistic, degrees of freedom, and critical value to determine if the observed distribution matches the expected one.

Output-

```
Chi-square Statistic: 22744.55
Degrees of Freedom: 4
Critical Value (alpha = 0.05): 9.49
```

```
Reject H0: gof_stat = 22744.55 > 9.49 ; the candidate probabilities do not fit the data.
```

Insights-

- Chi-Square Statistic (22744.55) > Critical Value (9.49), so we reject H_O
- The observed cut distribution does not fit the assumed uniform probabilities.

(c) T-Test for Mean

Consider one continuous variable in the dataset and apply test of mean for a proposed candidate of μ at the significant level alpha=0.05. (10)

Hypothesis -

The one-sample T-test checks whether the mean price of diamonds is significantly different from a proposed value (μ =4000).

Null Hypothesis (H_O): The mean price of diamonds is 4000 (μ = 4000).

Alternative Hypothesis (H_A): The mean price of diamonds is not 4000 ($\mu \neq 4000$)

Code -

```
candidate_mean = 4000
t_test_result = t.test(diamonds$price, mu = candidate_mean)

t_stat = t_test_result$statistic
df = t_test_result$parameter
critical_value = qt(0.975, df)

cat("T-statistic:", round(t_stat, 2),
    "\nDegrees of Freedom:", df,
    "\nCritical Value (alpha = 0.05):", round(critical_value, 2), "\n\n")

if (abs(t_stat) > critical_value) {
  cat("Reject H0: |t_stat| =", round(abs(t_stat), 2),
      ">", round(critical_value, 2),
      "; significant evidence that the mean is not", candidate_mean, "\n")
} else {
  cat("Fail to reject H0: |t_stat| =", round(abs(t_stat), 2),
      "<=", round(critical_value, 2),
      "; insufficient evidence to conclude the mean differs from",
      candidate_mean, "\n")
}
```

The code performs a one-sample T-test, comparing the observed mean price of diamonds to an assumed population mean of 4000. It extracts the T-statistic, degrees of freedom, and critical value.

Output-

```
T-statistic: -3.91
Degrees of Freedom: 53939
Critical Value (alpha = 0.05): 1.96
```

```
Reject H0: |t_stat| = 3.91 > 1.96 ; significant evidence that the mean is not 4000
```

Insights-

- | T-Statistic | (3.91) > Critical Value (1.96), so we reject H_O .
- Since T-statistic is negative, it means the sample mean price is lower than 4000.
- There is significant evidence that the mean diamond price is not 4000.