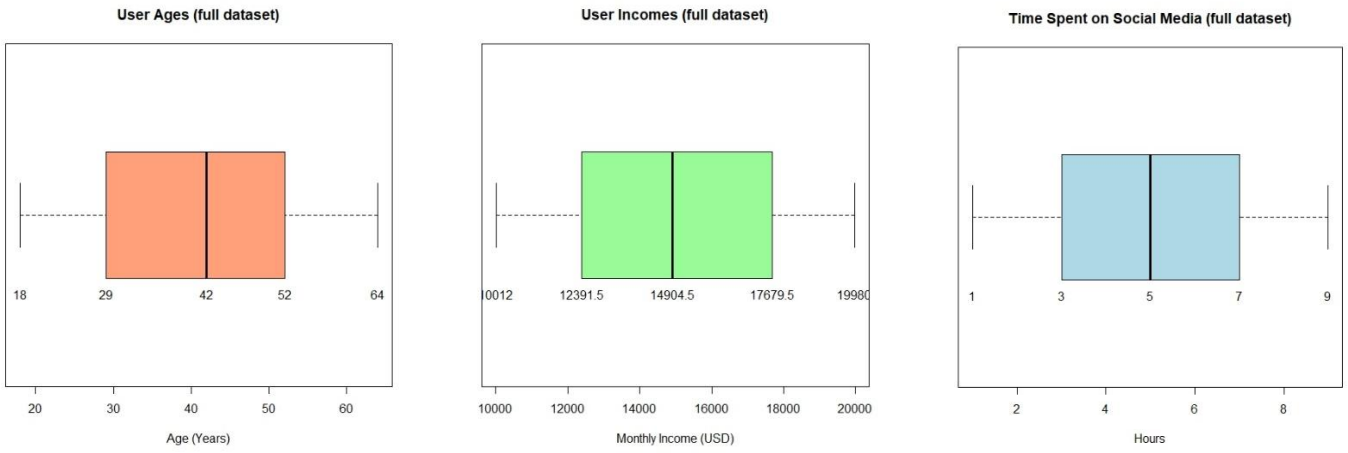
Social Media Data Analysis
Applied Statistics Final Project
Submitted by: Aditi Chadha & Anna Morin

Social Media Data Analysis

Introduction

For this statistical analysis, we examined a generated social media dataset[1] with the intention of discovering statistically significant answers to usage-, platform-, and interest-focused questions. The dataset consists of 1,000 entries containing user demographic information such as age, gender, interests, and social media platform utilized – (just to name a few). Our goals were to identify any factors affecting the average time spent on social media as well as identifying any statistically significant differences within the demographic groups listed above. We formed our analysis and testing processes to be repeatable. That is, for a different social media dataset, once cleaned and formatted to match the one used here, the tests performed in this analysis could easily be implemented again and different conclusions could potentially be formed.

The dataset contains dummy data generated to simulate a survey of 1000 users from varying countries, locations, demographics, and socioeconomic conditions. Since the data is randomly generated, we can create a highly representative sample that may be harder to recreate in real life, however, any conclusions that we draw cannot be extrapolated to the real world.



All three variables plotted above are discrete, symmetrically distributed, and contain no outliers. When creating models and visual aids that use these datasets, these qualities must be considered. We will also note that the variable "time_ spent" to describe the average time spent on social media does not have a defined unit, so for the purposes of this project we will assume the unit is hours per day.

[1] Link to raw data: https://drive.google.com/file/d/1z4hBQbak1zj-yXnwPZzU7t6BTvFnKdo3/view?usp=drive_link

Contents

Introduction 1
Usage-focused questions 3
Age and social media usage, model 3
Age and social media usage, model with buckets 4
Income and social media usage, model 5
Does gender affect the amount of time spent on social media? 6
Location and social media usage 6
Home/car ownership and social media usage 8
Demographics and social media usage 9
Platform-focused questions 10
Time spent on social media by platform 10
Median age of users by platform 10
Gender and platform 11
Location and platform 12
Platform and demographics 13
Interest-focused questions 14
Primary interest and gender 14
Primary interest and location 14
Proportion of users interested in Instagram, men and women 15
Conclusion 16

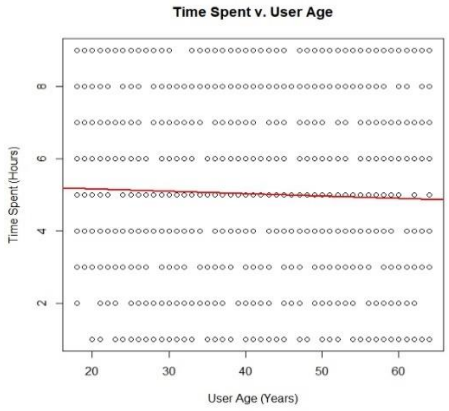
Usage-focused questions

Age and social media usage, model

We first consider whether a change in age has an effect on a user's average time spent on social media. We also hypothesize that variations in user age on the scale of one year may not be a useful experimental unit; as such, we later conduct an Analysis of Variance that group individuals by decade.

We will create a linear model using user age and average time spent on social media to test our null hypothesis that H_0: $\beta_1 = 0$ at the $\alpha = 0.05$ level.

Plotting the data to begin, it is not immediately obvious if there exists any correlation between the two variables; we recall the discrete nature of the dataset, causing many of our data points to overlap, which may hide the underlying trends.



Before creating the linear model, we evaluate the value of the Pearson correlation coefficient to understand what correlation (if any) exists in the sample data:

$$\hat{\rho} = \frac{\widehat{Cov}(Age, Time)}{s_{Age} s_{Time}} = \frac{-1.158753}{13.49785 \cdot 2.537834} = -0.03382696$$

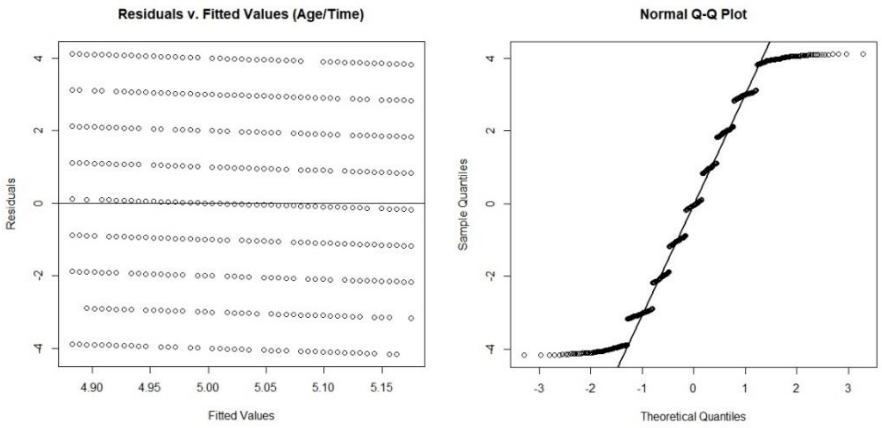
From this, we determine that there is a slight negative correlation between user age and average time spent on social media. Next, we will calculate the least squares coefficients of a linear model to test the null hypothesis $H_0: \beta_1 = 0$ to determine if there is a relationship between the two variables at the population level.

Using R, we output the summary of the linear model:

```
Call:
lm(formula = time_spent ~ age, data = social)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1625 -2.0734 -0.0448  2.0220  4.1174

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.289674   0.256660  20.610   <2e-16 ***
age         -0.006360   0.005948  -1.069    0.285
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.538 on 998 degrees of freedom
Multiple R-squared:  0.001144,  Adjusted R-squared:  0.0001434
F-statistic: 1.143 on 1 and 998 DF,  p-value: 0.2852
```

The p-value of the t-test is $0.285 > \alpha = 0.05$, which leads us to reject the null hypothesis and we conclude that user age and average time spent on social media are not related.

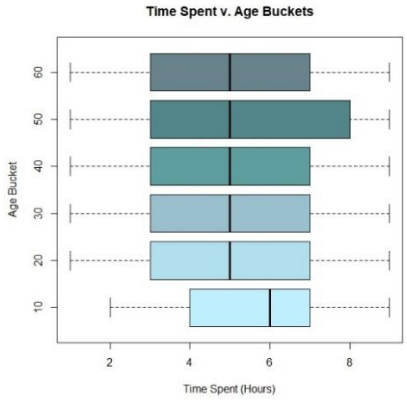
To confirm the validity of the regression, we plot the residuals against the fitted values and prepare a QQ plot to evaluate the variance and normality of the residuals.



We conclude that the variance of the residuals is symmetric, centered on zero, and apparently constant for all fitted values. Our QQ plot shows evidence of heavy-tailed behavior, but largely adheres to the normal distribution. We conclude that our model is not exhibiting any abnormal behavior that would call into question the validity of our earlier conclusion.

Age and social media usage, model with buckets

Before we evaluated the Age Model, we hypothesized that the difference between social media users of ages less than a year apart would be relatively insignificant, and grouping the age ranges by decade could be a more useful analysis. As such, we will run a one-way ANOVA on the bucketed data to see if the means differ between the six groups. First, we plot the different buckets by decade to understand what trends we can expect.



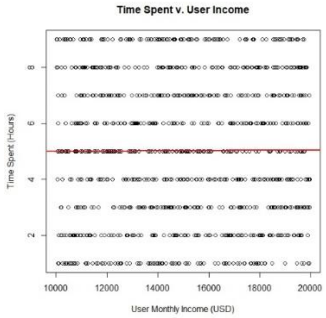
From this graph, it appears that most groupings are distributed very similarly, which may suggest that the group means are all identical. The null hypothesis H_0: (all means are the same) is tested at the $\alpha = 0.05$ level in R, giving us the following output:

```
Df Sum Sq Mean Sq F value Pr(>F)
age.bucket    5     26   5.206   0.808  0.544
Residuals   994   6408   6.447
```

The F-test results in a p-value of 0.544, and we fail to reject the null hypothesis at the $\alpha = 0.05$ level. As a result, we conclude that the group population means are the same.

Income and social media usage, model

We now consider that income may be related to the average time spent on social media - it is possible that a wealthier user would have more spare time to spend on social media. Initially, the scatter plot appears to show no trend; however, since "Time Spent" is a discrete variable, it is harder to identify underlying patterns.



Before creating the linear model, we again evaluate the value of the Pearson correlation coefficient to understand what correlation (if any) exists in the sample data:

$$\hat{\rho} = \frac{\widehat{Cov}(Income, Time)}{s_{Income} s_{Time}} = \frac{35.71985}{2958.628 \cdot 2.537834} = 0.004757252$$

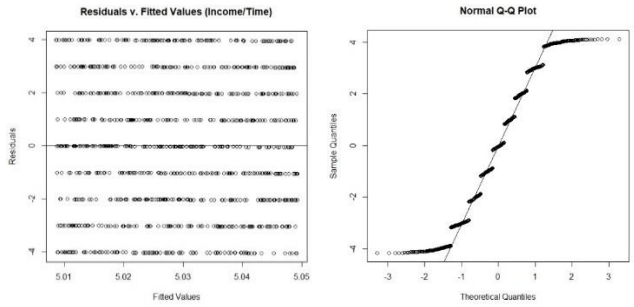
From this, we determine that there is a slight positive correlation between user age and average time spent on social media. However, the value is very close to zero, which indicates there may be no correlation.

Next, we test the null hypothesis $H_0: \beta_1 = 0$ to determine if there is a relationship between the two variables at the population level. Using R, we output the summary of the linear model:

```
Call:
lm(formula = time_spent ~ income, data = social)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0492 -2.0372 -0.0258  1.9796  3.9914

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.968e+00  4.155e-01   11.96   <2e-16 ***
income      4.081e-06  2.715e-05    0.15    0.881
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.539 on 998 degrees of freedom
Multiple R-squared:  2.263e-05, Adjusted R-squared:  -0.0009793
F-statistic: 0.02259 on 1 and 998 DF,  p-value: 0.8806
```

The p-value of the t-test is $0.881 > \alpha = 0.05$, which leads us to reject the null hypothesis and we conclude that user monthly income and average time spent on social media are not related.

To confirm the validity of the regression, we plot the residuals against the fitted values and prepare a QQ plot to evaluate the variance and normality of the residuals.



Again, we conclude that the variance of the residuals is symmetric, centered on zero, and constant for all fitted values. The QQ plot is similarly heavy-tailed to our prior age model, but largely adheres to the normal distribution. We conclude that our linear model is not exhibiting any abnormal behavior that would call into question the validity of our conclusion.

Does gender affect the amount of time spent on social media?

In this data set we have three genders, female, male and non-binary and we assumed that the unit of measurement for the variable labeled 'time_spent' is hours per day. We want to conduct a one-way ANOVA test to see if the mean amount of time spent on social media is the same for all genders included in the study. To use the one-way ANOVA test we had to make some assumptions. These assumptions include: the data being normally distributed, the variance of the data within each group is equal, and the observations within each group are independent. The hypotheses we are testing are:

Ho: $\mu_{Female} = \mu_{Male} = \mu_{Non\text{-}Binary}$ vs. H_A: At least one mean is different

The rejection rule is to reject the null hypothesis if $F_{observed} > F_{k-1,\ n-k,\ \alpha}$. For this test, we decided to use a significance level of 0.05. We have three groups (k=3) and 1000 data entries (n=1000).

```
> ## One-Way Anova
> summary(aov(time_spent ~ gender, data=final.data))
             Df Sum Sq Mean Sq F value Pr(>F)
gender        2     24   11.93   1.856  0.157
Residuals   997   6410    6.43
```

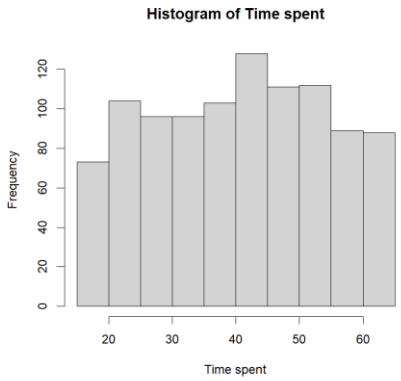
Looking at the ANOVA table, we can see that the p-value is 0.157, which is greater than our α level of significance. Thus, we fail to reject the null hypothesis at the 0.05 significance level and fail to conclude any different mean time spent values among the three genders.

Location and social media usage

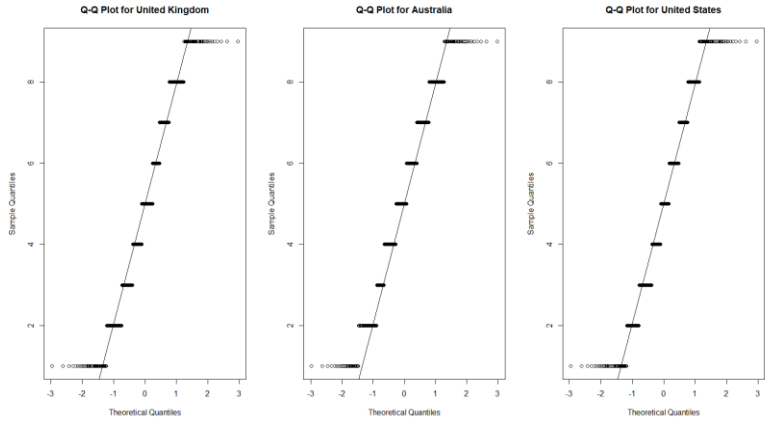
The dataset includes data collected from three unique geographical locations, each representing a different user base utilizing social media platforms. The objective of this inquiry is to investigate whether geographical location influences user behavior in terms of time spent on social media. Initially, we sought to assess the potential discriminatory influence of location on social media usage patterns. In other words, we aimed to determine if particular locations demonstrate discernible variations in social media engagement.

From the above chart, we can observe that the box plots do not offer any clear separation of discriminatory power between the three locations. The range of time spent on social media is similar across all three countries. We may concur that, despite cultural differences and geographical distances, people in these countries exhibit a consistent pattern of social media usage and it appears that location alone may not significantly impact social media usage.

As a next step, we wish to determine ***if the population means of the three countries are identical for time spent on social media***. Prior to proceeding, we aimed to ascertain the most appropriate statistical test—whether parametric or non-parametric. We conducted an analysis of the Time spent variable's histogram to evaluate the distribution's conformity to a normal curve.



A normal distribution is symmetric, with equal probabilities on both sides of the mean. In this histogram, the data is skewed to the right (positively skewed). The tail on the right side is longer, indicating that some individuals spend significantly more time on social media. In order to further confirm if the "time_spent" column is normal, we explored the Q-Q plots.



A Q-Q plot helps us assess whether a dataset follows a specific theoretical distribution (such as a normal distribution). The x-axis represents theoretical quantiles (expected data distribution), and the y-axis represents sample quantiles (actual data distribution). Based on the Q-Q plots, we cannot confidently say that the data is perfectly normal. However, the deviations are not extreme, suggesting that the data may approximately follow a normal distribution. Hence, we would continue to use a parametric test – specifically, a one way analysis of Variance to compare the sample means across the three countries.

H_0: The Population means of the three countries for time spent on social media are identical

H_a: At least one of the population means differs from one of the other countries

The one –way ANOVA test yields the following result:

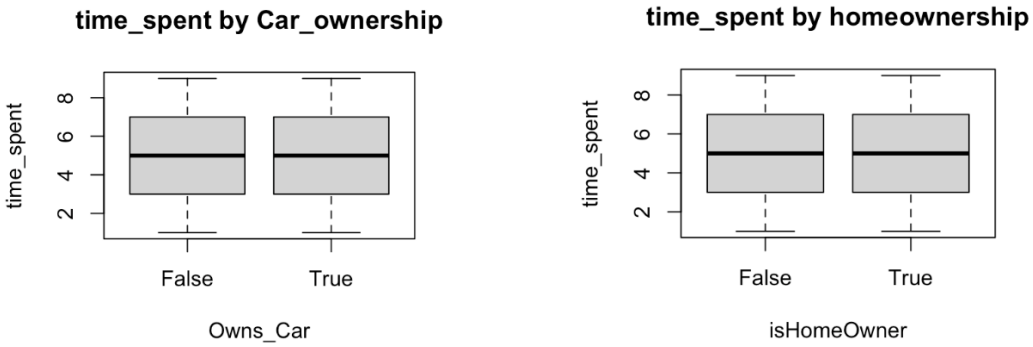
```
>                                                    summary(res.aov)
             Df      Sum      Sq      Mean      Sq      F      value      Pr(>F)
location                2                  20          9.877          1.535      0.216
Residuals    997    6414    6.434
```

Since the p-value (0.216) is greater than the common significance level (α = 0.05), we fail to reject the null hypothesis. This means there isn't enough evidence to conclude that there is a significant difference in the population means of time spent on social media among the three countries. **We find no significant evidence to suggest that the population**

means of time spent on social media differ across different locations. Therefore, location does not appear to impact social media usage.

Home/car ownership and social media usage

The variables that we have for home/car ownership are binary, i.e. "True" signifying ownership and "False" signifying no ownership. For every individual, we also have their respective time spent on social media in hours (continuous variable). Since the aim of this section is to investigate the potential impact of home and car ownership on social media usage, we start with plotting boxplots to check if there are any visually significant differences based on groups of ownership.



Looking at the above plots, we tend to infer that there is almost no difference in the distribution of time_spent between groups of ownership, indicating independence of time_spent with respect to house or car ownership. To further investigate this, and the possibility of a combined interactive effect on the time_spent, we proceed by conducting hypothesis tests. Although, it might occur intuitively to go with a t-test (based on variable's nature), a two-way ANOVA assuming normality within groups and equal variances, would be a better test to check if there is any significant relationship between time_spent and ownership in two variables, especially if we also want to consider their interactive effect.

With a Two-way ANOVA Test, we consider the following hypothesis:

- (H0): There is no difference in the mean time spent on social media between homeowners and non-homeowners.
- (H0): There is no difference in the mean time spent on social media between car owners and non-car owners.
- (H0): There is no interaction effect between homeownership and car ownership on time spent on social media.

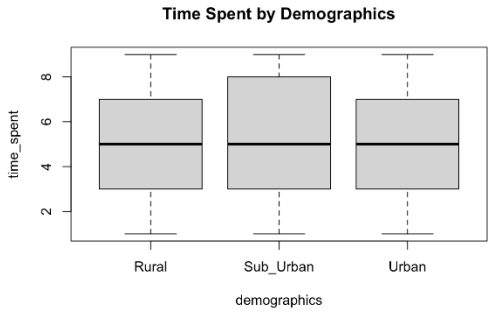
Results:

```
> summary(two_way_anova)
                     Df Sum Sq Mean Sq F value Pr(>F)
isHomeOwner           1      6   5.557   0.861  0.354
Owns_Car              1      2   2.270   0.352  0.553
isHomeOwner:Owns_Car  1      1   0.892   0.138  0.710
Residuals           996   6425   6.451
```

The p-value associated with the factor isHomeOwner is 0.3540, the p-value associated with the factor Owns_Car is 0.5530, and the p-value associated with the interaction term isHomeOwner:Owns_Car is 0.7100. The p-values associated with all our hypothesis tests are greater than 0.05. Therefore, at a significance level of 0.05, we fail to reject all the null hypotheses, as there is not enough significant evidence to prove otherwise. This confirms our initial inference of significant independence between time_spent, ownership variables and their interaction. Based on the results of the two-way ANOVA, neither homeownership nor car ownership, nor their interaction, have a statistically significant effect on the time spent on social media within the context of the dataset analyzed.

Demographics and social media usage

To examine any clear visual differences in distributions between groups of demographics, we plot a box plot.



The boxplot indicates a difference in distribution of time_spent for the Sub_Urban demographic with a similar central value, compared to other groups. This might either indicate a presence of few extreme values in the Suburban group or possibly, a dependence between the time_spent on social media and demographics. To further investigate this, assuming Normality within groups and equal variances, we proceed with a one-way ANOVA test with a null hypothesis:

(H0): The mean time spent on social media is the same across all demographic groups.

```
> summary(anova_result)
              Df Sum Sq Mean Sq F value Pr(>F)
demographics   2     39  19.336   3.014 0.0495 *
Residuals    997   6395   6.415
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

This p-value is very close to the common significance level of 0.05. It indicates that there is a statistically significant difference in the mean time spent on social media among the different demographic groups, but just barely. This result suggests that demographics do influence how much time individuals spend on social media, though the effect might be small given the proximity of the p-value to 0.05.

Given the borderline p-value, it might be insightful to conduct post-hoc tests (e.g., Tukey's HSD) to determine which specific groups differ from each other.

Tukey’s HSD test

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = time_spent ~ demographics, data = data)

$demographics
                     diff        lwr        upr     p adj
Sub_Urban-Rural  0.2510536 -0.2065946  0.7087017 0.4024784
Urban-Rural     -0.2328959 -0.6940769  0.2282850 0.4621915
Urban-Sub_Urban -0.4839495 -0.9468094 -0.0210896 0.0379679
```

From this output, observe:

- The mean time spent on social media by Sub_Urban is not significantly different from Rural (p adj = 0.4024784).
- There is also no significant difference between the Urban and Rural groups in terms of time spent on social media (p adj = 0.4621915).
- The difference between Urban and Sub_Urban is statistically significant with Urban spending less time on social media compared to Sub_Urban (p adj = 0.0379679).

The only statistically significant difference found is between Urban and Sub_Urban, where the adjusted p-value is below the common alpha level of 0.05. This suggests that there are indeed some differences in how much time is spent on social media across different demographic groups, particularly between the Urban and Sub_Urban categories.

Platform-focused questions

Time spent on social media by platform

In examining the time spent on social media, we are interested in determining if this time is significantly different across the three platform types listed. Using R, we first examine the summary statistics of time spent by social media platform.

```
 Descriptive statistics by group
group: Facebook
   vars   n mean  sd median trimmed  mad min max range skew kurtosis   se
X1    1 307 5.06 2.5      5    5.05 2.97   1   9     8 0.01    -1.19 0.14
------------------------------------------------------------------------
group: Instagram
   vars   n mean   sd median trimmed  mad min max range  skew kurtosis   se
X1    1 363 5.15 2.61      5    5.19 2.97   1   9     8 -0.03    -1.24 0.14
------------------------------------------------------------------------
group: YouTube
   vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 330 4.87 2.49      5    4.85 2.97   1   9     8 0.04    -1.14 0.14
```

We observe that the mean values are all different, and we are interested in testing the null hypothesis that the mean time spent is the same across all platforms. The corresponding alternative hypothesis is that at least one mean is different from the others and overall population mean. To test the null hypothesis at the $\alpha = 0.05$ significance level, we consider a one-way ANOVA table (below). The p-value for this test is $0.337 > 0.05$, and we therefore fail to reject the null hypothesis. As a result, we fail to find any evidence of different amounts of time spent on different platforms.

```
> smd$platform=as.factor(smd$platform)
> smd.fit = aov(time_spent~platform, data = smd)
> summary(smd.fit)
             Df Sum Sq Mean Sq F value Pr(>F)
platform      2     14   7.018    1.09  0.337
Residuals   997   6420   6.439
```

We now wish to consider pairwise comparisons between the different platforms using Tukey HSD and LSD tests. In these, we test the null hypothesis that the mean time spent between each of the potential pairings are equal. The corresponding alternative hypothesis is that the two means are different for a given pairing. Using R, the output for these tests is below.

```
> #tukey HSD
> TukeyHSD(smd.fit, confidence.level=0.95)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = time_spent ~ platform, data = smd)

$platform
                         diff        lwr       upr     p adj
Instagram-Facebook  0.09614056 -0.3656993 0.5579804 0.8767410
YouTube-Facebook   -0.18567762 -0.6579799 0.2866247 0.6260038
YouTube-Instagram  -0.28181818 -0.7348543 0.1712179 0.3107138
```

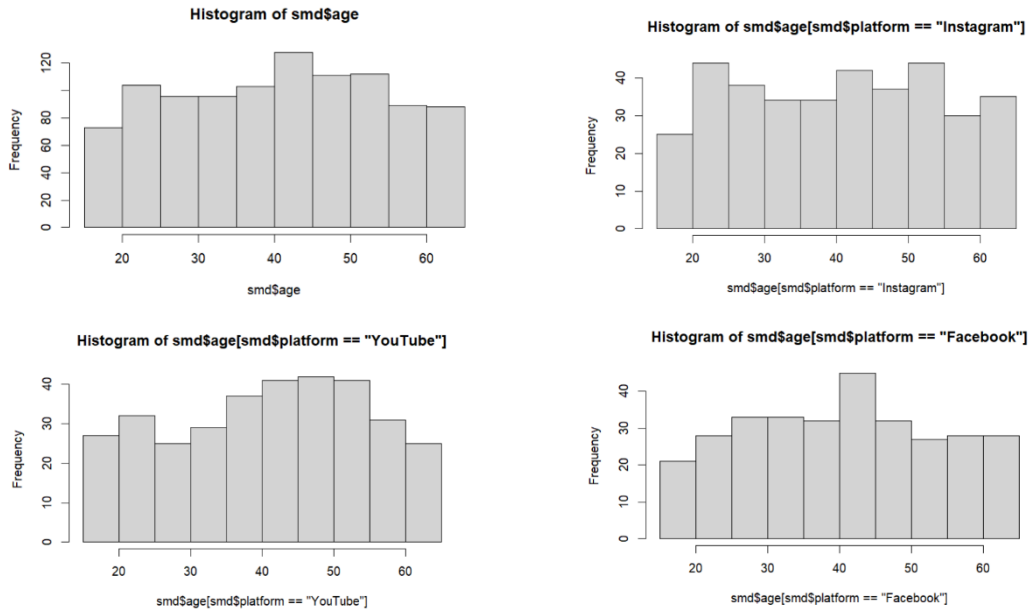
```
> #Tukey LSD:
> pairwise.t.test(smd$time_spent,g=smd$platform,p.adjust.method='none')

        Pairwise comparisons using t tests with pooled SD

data:  smd$time_spent and smd$platform

          Facebook Instagram
Instagram 0.63     -
YouTube   0.36     0.14

P value adjustment method: none
```

In the Tukey HSD output, the p-values for Instagram-Facebook, YouTube-Facebook, and YouTube-Instagram and 0.8767, 0.6260, and 0.3107 respectively. In the Tukey LSD output, the pairwise comparison test p-values are 0.63, 0.36, and 0.14 (following the same ordering). For both the HSD and LSD comparison tests, every p-value is greater than $\alpha = 0.05$, and we fail to conclude that any two pairs have significantly different mean values of time spent on social media. We do note that the YouTube-Instagram test p-value is the lowest across all pairwise tests in both the Tukey HSD and LSD tests. While this value is still larger than 0.05 and thus not significant, it signals that the YouTube-Instagram means are more likely to be differentiated than the other pairs. This is observable in the summary statistics on the previous page, as the Instagram and YouTube means of 5.15 and 4.87 are the farthest apart of any two means by platform.

Median age of users by platform

In examining the age of people on social media, we are interested in determining if the median ages of users are different across platforms. A cursory examination of summary statistics output does not suggest that the medians would be

different, but to be certain we will formalize a hypothesis test. Before determining the appropriate test, we will examine the histogram age outputs for the full sample and the three different platforms.

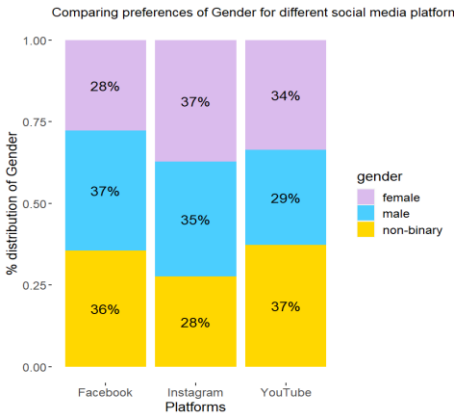


These histograms, representing the age distribution by the whole sample and the three different platforms, are different enough to warrant the use of the Kruskal Wallis test. We wish to test the null hypothesis that the median age of the user is the same across all platforms. The corresponding alternative hypothesis is that at least one median age value is different from the others. Using the Kruskal Wallis test and the $\alpha = 0.05$ significance level, we obtain a p-value of 0.9203 and fail to reject the null hypothesis. We fail to conclude that there is any difference in median age across the platforms. The R output for the test is below:

```
> kruskal.test(smd$age,smd$platform)

        Kruskal-Wallis rank sum test

data:  smd$age and smd$platform
Kruskal-Wallis chi-squared = 0.16608, df = 2, p-value = 0.9203
```

Gender and platform

Here, our objective is to investigate whether there exists a statistically significant preference of any gender group for a specific social media platform type. Using 100% stacked graph, we visualized the proportions of gender across different platforms by displaying the distribution of each gender category as a percentage of the total for each platform.



From the above graph, we observe that social media platform preferences vary by gender, with distinct patterns across the three platforms. Females show a higher preference for Instagram and YouTube compared to Facebook. Males

tend to favor Facebook and Instagram over YouTube. Non-binary individuals lean toward Facebook and YouTube, with less preference for Instagram. In order to further investigate this we performed a chi-squared test with below hypotheses:

H_0 : Within each platform types, the proportions of gender types are identical

H_a: The proportions of gender within each platform types are not identical

Result:

	female	male	non-binary
Facebook	85	135	111
Instagram	113	128	96
YouTube	109	100	123

```
>                                                    chi_sq_result
      Pearson's                      Chi-squared                  test

data:                                                contingency_table
X-squared = 13.436, df = 4, p-value = 0.009331
```

With a p-value of 0.009331, which is less than a typical significance level of 0.05, there is sufficient evidence to **reject the null hypothesis (H0)** that the platform type and gender variables are independent. The significant p-value suggests that the choice of social media platform may not be independent of an individual's gender.

While the chi-squared test tells us that there is a significant association between the variables, it doesn't provide information about the strength or magnitude of that association. We conducted Cramer's V test to confirm the strength of the association:

```
>             #             Print             the             result
>          cat("Cramer's          V:          ",          cramers_v)
Cramer's V:  0.08196317
```

A Cramer's V of 0.082 (with degrees of freedom = 1) indicates a small (or "weak") association between gender and social media platform preference. This means that **while there is an association, it is not very strong.**

Location and platform

Here, the variables we are dealing with are both categorical. Thus, to evaluate their relationship, we shall conduct a Chi-square test of independence. To get a better understanding of the observed frequencies in each combination of groups, we create a 3x3 contingency table.

Platform	Australia	United Kingdom	United States	Total
Facebook	106	107	94	307
Instagram	125	117	121	363
YouTube	121	105	104	330
Total	352	329	319	1000

This table shows the observed frequencies for each category combination, the total counts per platform, total counts per location, and the grand total of observations. This comprehensive view is essential for conducting the Chi-square test manually, as it helps in calculating the expected frequencies under the null hypothesis of independence.

In carrying out the test, (H0): the choice of primary platform is independent of the user's location, the first step is to calculate the expected count for each cell in the contingency table.

The expected count is equal to the row total multiplied by the column total divided by the table total, (expected count table below).

Platform	Australia	United Kingdom	United States	Total
Facebook	108.064	101.003	97.933	307.0
Instagram	127.776	119.427	115.797	363.0
YouTube	116.160	108.570	105.270	330.0
Total	352.0	329.0	319.0	1000.0

Now, for the chi-square statistic: $X^2 = \sum_{i=1}^{9} \frac{(Oi-Ei)^2}{Ei}$, where O_{ij} represents the observed frequency in the i'th row and j'th column of our contingency table. E_{ij} represents the expected frequency for the i'th row and j'th column, calculated under the null hypothesis of independence.

We get a chi-square value of approximately 1.2312, referring this along with a degree of freedom (df) = (3-1) *(3-1) = 4 to a Chi-square distribution table, we get a corresponding p-value of 0.8729. The p-value associated with our hypothesis test is much greater than 0.05, i.e. at a significance level of 95% (0.05), we fail to reject the null hypothesis and conclude that primary platform is independent of the user's location.

Platform and demographics

We now wish to check the dependence between the primary platform and the demographics, we follow a similar procedure as the one used to check dependence between primary platform and location.

3x3 contingency table of observed frequencies

Platform	Rural	Sub-Urban	Urban	Total
Facebook	96	106	105	307
Instagram	136	118	109	363
YouTube	108	111	111	330
Total	340	335	325	1000

3x3 contingency table of expected frequencies

Platform	Rural	Sub-Urban	Urban	Total
Facebook	104.38	102.845	99.775	307.0
Instagram	123.42	121.605	117.975	363.0
YouTube	112.20	110.550	107.250	330.0
Total	340.0	335.0	325.0	1000.0

Chi-square Statistic: $\chi 2 = 3.405$, Degrees of Freedom: 4, P-value: 0.492

The p-value is significantly higher than 0.05. This indicates that we fail to reject the null hypothesis at the 95% level of significance, suggesting there is no sufficient evidence to conclude that demographics and primary platform are dependent.

Interest-focused questions

Primary interest and gender

We wish to determine if there is any association between primary interest and gender. To test for a potential association, we first construct a 3x3 contingency table. We also construct a 3x3 table of expected values obtained by taking the product of the corresponding row and column sums for each table location and dividing by the total sum. The 3x3 contingency table and table of expected values are included below on the left and right respectively.

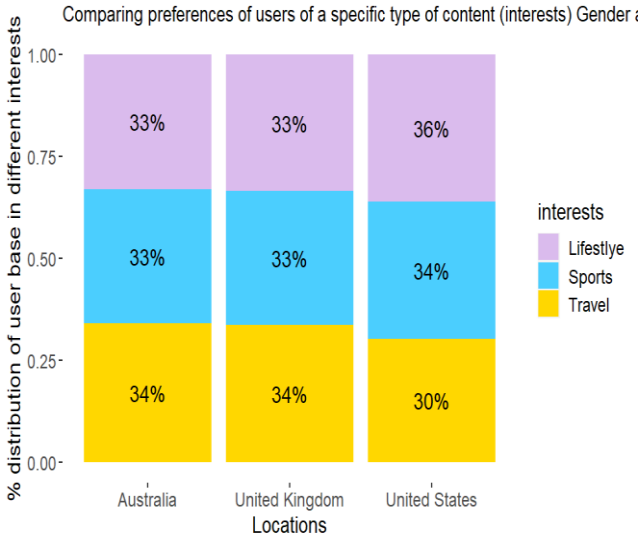
Gender	Primary interest			Total
	Sports	Travel	Lifestlye	
male	110	119	108	337
female	109	100	122	331
non-binary	112	109	111	332
Total	331	328	341	1000

Gender	Primary interest			Total
	Sports	Travel	Lifestlye	
male	111.547	110.536	114.917	337
female	109.561	108.568	112.871	331
non-binary	109.892	108.896	113.212	332
Total	331	328	341	1000

We wish to test the null hypothesis that there is no association between gender and primary interest at the $\alpha = 0.05$ significance level. The alternative hypothesis is that there is an association between gender and primary interest. We let Oi and Ei represent the corresponding values from the observed and expected contingency tables respectively and calculate the test statistic $X^2 = \sum_{i=1}^{9} \frac{(Oi-Ei)^2}{Ei} = 2.5871$. Comparing this to 9.4877, (Chi square value for $\alpha = 0.05$ and (3-1)(3-1) = 4 degrees of freedom), we see that $2.5871 < 9.4877$ and fail to reject the null hypothesis. The p-value for this test is 0.6291, and we do not conclude that there is any association between primary interest and gender.

Primary interest and location

Here, our objective is to investigate if there exists a statistically significantly preference of content type (interests) in different countries (location). Using 100% stacked graph, we visualized the proportions of users having specific interests across different locations by displaying the distribution of each interest's category as a percentage of the total for each location.



From the above graph, we can observe that across Australia and United Kingdom, there exists equal proportion of users who are interested in using social media for Lifestyle, Sports and Travel whereas in United States, more users prefer to use the media for Lifestyle and Sports related interests.

In order to further investigate this, we performed a chi-squared test with below hypotheses:

H_0 : Within each location, the proportions of users having different interests are identical
H_a: The proportions of users having different interests are not identical across different locations.

	Australia	United Kingdom	United States
Lifestyle	116	116	120
Sports	110	108	111
Travel	115	107	97

```
        Pearson's Chi-squared test
data:  data_df$interests and data_df$location
X-squared = 1.3863, df = 4, p-value = 0.8466
```

The p-value obtained from the chi-squared test is 0.8466. Since this p-value is greater than 0.05, we do not have enough evidence to reject the null hypothesis. Hence, there is no significant difference in the proportions of users with different interests across different locations based on the data analyzed.

Proportion of users interested in Instagram, men and women

One final platform-specific question is if the proportion of men and the proportion of women interested in Instagram are the same. We have a total of 668 individuals in the data set who identify as male or female. Of those 668 people, 331 of them are female and 337 are male.

Our hypothesis for this test is:

Ho: $p_{Female} = p_{Male} = p$ vs. H_A: $p_{Female} \neq p_{Male}$

Under the null hypothesis:

$$X_{Female} \sim Bin(n_1, p), \quad X_{Male} \sim Bin(n_2, p), \quad X_{Female} + X_{Male} \sim Bin(n_1 + n_2, p)$$

Since p is unknown so we estimate by pooling, $\hat{p} = \frac{x_{female}+x_{male}}{n_{female}+n_{male}} = \frac{134+137}{330+336} = 0.392$

We will use $\alpha = 0.05$ level of significance for this test. The rejection rule is to reject the null hypothesis if $z_{obs} > z_{1-\frac{\alpha}{2}}$.

The test statistic is $= \left| \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(\hat{p})(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \right| = \left| \frac{0.406 - 0.378}{\sqrt{0.392(1-0.392)\left(\frac{1}{330}+\frac{1}{336}\right)}} \right| = 12.282 > z_{1-\frac{\alpha}{2}} = z_{1-\frac{0.05}{2}} = 1.96$

Since, $z_{obs} > z_{1-\frac{\alpha}{2}}$, we reject the null hypothesis at the 0.05 level of significance and conclude that the proportion of females who use Instagram is not equal to the proportion of males who use Instagram. Below is the R-code used to conduct this test:

```
> ## H0: The proportion of users interested in Instagram is the same between men/women
> # Break the data up into female and male variables. Subtract one for the column names.
> female = (subset(final.data, gender == 'female'))
> (num.female = nrow(female)-1)
[1] 330
> male = (subset(final.data, gender == 'male'))
> (num.male = nrow(male)-1)
[1] 336
> # Within male and female, calculate the number of people who use Instagram
> female.Instagram = subset(female, platform == 'Instagram')
> (num.Female.Instagram = nrow(female.Instagram)-1)
[1] 134
> male.Instagram = subset(male, platform == 'Instagram')
> (num.male.Instagram = nrow(male.Instagram)-1)
[1] 127
> # proportion of female and proportion of male who use instagram
> (prop.female = num.Female.Instagram/num.female)
[1] 0.4060606
> (prop.male = num.male.Instagram/num.male)
[1] 0.3779762
> # estimated p
> (est.p = (num.Female.Instagram+ num.male.Instagram)/(num.female+num.male))
[1] 0.3918919
> # z-obs
> (z.obs = abs((prop.female-prop.male)/(sqrt(est.p)*(1-est.p)*(1/num.female+1/num.male))))
[1] 12.28232
> qnorm(1-(0.05/2))
[1] 1.959964
```

Conclusion

For this statistical analysis, we examined a generated social media dataset with the intention of discovering statistically significant answers to usage-, platform-, and interest-focused questions. We developed regression models to model the time spent on social media dependent on age and income, and we performed numerous hypothesis test to examine other key questions within our three main focus areas. Although we were not able to reject most of our null hypotheses at the 0.05 level, we managed to form statistical conclusions where applicable and carried out tests we deemed appropriate for every question we asked.

In the usage-focused section, we concluded that differences in usage time by Urban and Sub_Urban demographic segments is statistically significant with Urban spending less time on social media compared to Sub_Urban. In the platform-focused section, we determined that there is an association between gender and social media preference. Further, in the interest-focused section, we concluded that the proportion of men and women using Instagram was not equal. Our process developed within this analysis is repeatable on any similarly formatted dataset, and we believe the methods used to obtain our results were reasonable and appropriate. Before carrying out this analysis on another dataset, it would be prudent to again check that assumptions embedded in the tests we used are not potentially violated.

[*] Indicates joint effort.