ST 314 Data Analysis 6

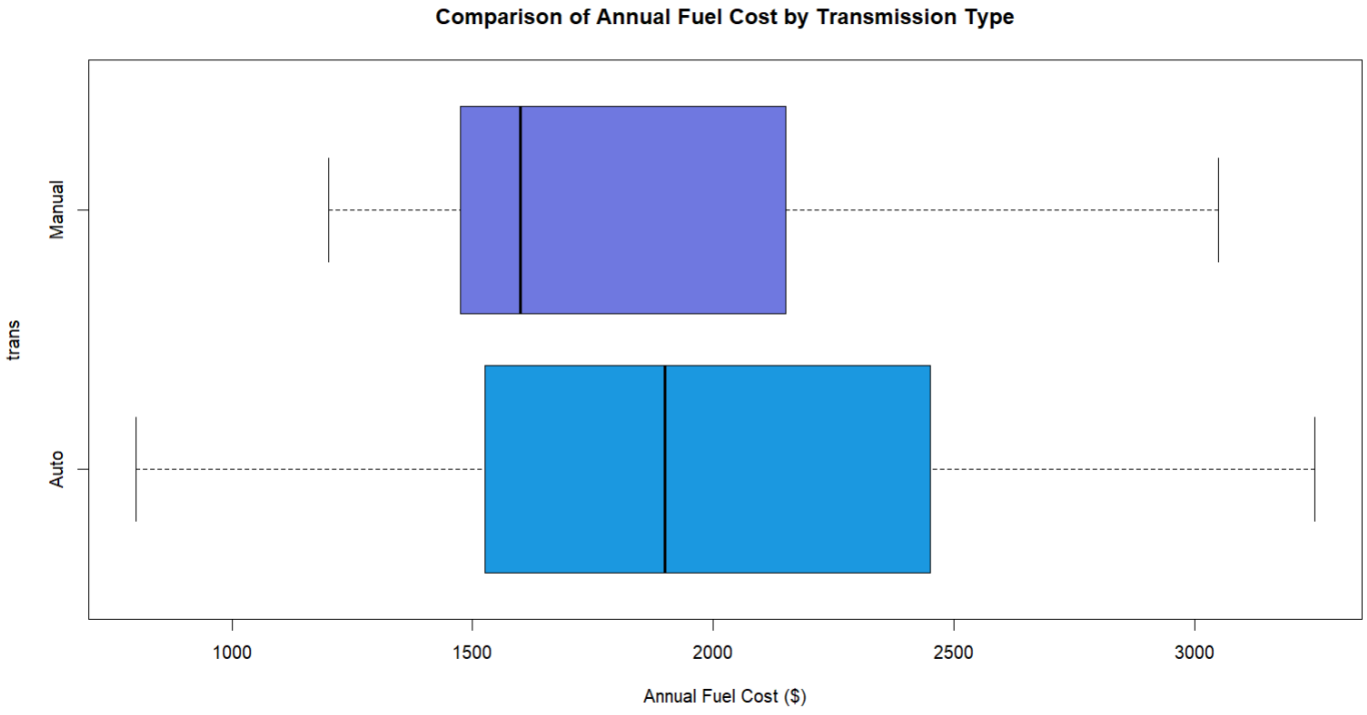
The following information will be used for parts 1 and 2 of this assignment.

The EPA_W24.csv contains a set of randomly sampled vehicles. Each vehicle in the data set has a projected Annual Fuel Cost that estimates how much the vehicle will cost over a year in fuel, along with many other variables about the vehicle.

Use the EPA_W24.csv dataset and the **ComparingTwoMeans_DA6.R** script to compare the average annual fuel cost in dollars between automatic and manual transmissions from 2021 vehicles.

Part 1. (6 points) Exploring the data.

a. (2 points) Visualize the data with a side by side boxplot.
   i. (1 point) Paste the side-by-side box plot of the data.
   ii. (1 point) Add a title and axis labels.
   iii. Optional to change the color of the plot.



b. (2 points) Describe the distribution in context. Is there visual evidence the average annual fuel cost is different between the automatic and manual vehicles? Explain.
The median annual fuel cost for manual vehicles is visibly lower than that of automatic vehicles. The medians are represented by the lines inside the boxes, and there is a clear gap between the two, with the median for automatic vehicles being to the right compared to that of manual vehicles.

c. (2 points) Provide an organized table of the summary statistics. Include the sample means, standard deviations and sample sizes for each group. Round to nearest whole number.

	MEAN	STANDARD DEVIATION	SAMPLE SIZE
AUTO	1983	615	35
MANUAL	1776	449	35

Part 2. (8.5 Points) Hypothesis Testing and Estimation.

Question of interest: Do the data provide evidence of a difference between the average annual fuel cost for automatic and manual transmissions?

Hypotheses: The null and alternative hypotheses are as follows, where μ_A represents the average annual fuel cost for *all* vehicles with automatic transmissions and μ_M represents the average annual fuel cost for *all* vehicles with manual transmissions.

$$H_0: \mu_A - \mu_M = 0$$
$$H_A: \mu_A - \mu_M \neq 0$$

Checking Conditions:

- The data available come from random samples from the population of all vehicles manufactured in 2021, so we can assume the sample are representative of their respective populations.
- The sample sizes are large enough so that the sampling distributions for $\bar{X}_A$ and $\bar{X}_M$ are both normal according to the central limit theorem.
- Lastly, the populations are independent. There is no repeated measurement nor is there any dependence between the two groups.

Overall, the conditions are somewhat met. The sampling method is unknown so we should consider this in our conclusions.

Calculate:

a. (1 point) From the summary statistics calculate the test statistic "by hand". *Show work.* You must show how the calculation for the test statistic is done (it is not enough to just show R output for this question).

$$t = \frac{(\overline{X_A} - \overline{X_M}) - (\mu_A - \mu_M)}{\sqrt{\left(\frac{s_A^2}{n_A}\right) + \left(\frac{s_M^2}{n_M}\right)}}$$

$$t = \frac{1983 - 1776}{\sqrt{\left(\frac{615^2}{35}\right) + \left(\frac{449^2}{35}\right)}}$$

$$t = \frac{1983 - 1776}{128.73} = \frac{207}{128.73} \approx 1.61$$

b. (1 point) State the degrees of freedom. You may choose conservative or Satterthwaite. Either are okay.

df = 62.25

c. (1 point) Obtain a p-value based on your calculated test statistic and degrees of freedom using the pt() function in R. Show work/code.

p-val = 0.1125

```
tVal <- 1.61

df <- 62.25

p_value <- 2 * pt(-abs(tVal), df)
p_value
```

d. (2 points) From the summary statistics, calculate the 95% Confidence Interval "by hand". Show work.

$$\text{SE} = \sqrt{\left(\frac{615^2}{35}\right) + \left(\frac{449^2}{35}\right)} = \sqrt{10806.43 + 5761.46} = \sqrt{16567.89} \approx 128.73$$

$$ME = t_{\alpha/2} \times \text{SE} = 2.0003 \times 128.73 \approx 257.57$$

$$(1982.857 - 1775.714) \pm 257.57$$

$$[-50.427, 464.287]$$

e. (0.5 point) Obtain a p-value from t-test and confidence interval using R. Paste the output. Are your answers different? Why, yes/no?

```
95 percent confidence interval:
-50.04407 464.32978
```

The manually calculated confidence interval and the confidence interval from the R output are very close, with only minor differences due to rounding. The p-values are the same in both cases: 0.1125.

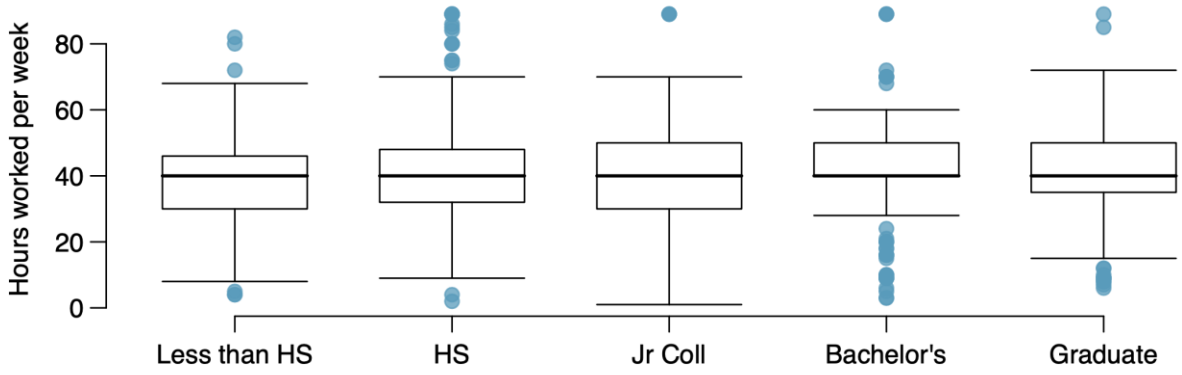
Conclude:

f. From the R output, write a four-part conclusion describing the results.
   - (1 point) Provide a statement in terms of the alternative hypothesis.
   - (1 point) State whether (or not) to reject the null.
   - (1 point) Give in context an interpretation of the point and interval estimate.
     - Make sure to provide a *direction* to your interval, for example, one group had a smaller (or larger) mean than the other, include this relationship in your point and interval estimate.

Based on the two-sample t-test, the p-value of 0.1125 indicates that we do not reject the null hypothesis, as it is greater than the 0.05 significance level. This means there is no significant difference in the average annual fuel cost between automatic and manual vehicles. The sample mean show that automatic vehicles have an average annual fuel cost of $1982.86, which is $207.14 higher than the $1775.71 for manual vehicles. However, the 95% confidence interval for the difference in means $[-50.04, 464.33]$ includes zero, suggesting the true difference could range from $50.04 less to $464.33 more for automatic vehicles, meaning the difference is not statistically significant.

Part 3. (8.5 points) ANOVA

The General Social Survey collects data on demographics, education, and work, among many other characteristics of US residents. Using ANOVA, we can consider educational attainment levels for all 1,172 respondents at once. Below are the distributions of hours worked by educational attainment and relevant summary statistics that will be helpful in carrying out this analysis.



	Educational attainment					
	Less than HS	HS	Jr Coll	Bachelor's	Graduate	Total
Mean	38.67	39.6	41.39	42.55	40.85	40.45
SD	15.81	14.97	18.1	13.62	15.51	15.17
n	121	546	97	253	155	1,172

a. (2 points) Write the null and alternative hypotheses for evaluating whether the average number of hours worked varies across the five groups.
There is no difference between average hours worked per week

$$\mu_{\text{Less than HS}} = \mu_{\text{HS}} = \mu_{\text{Jr Coll}} = \mu_{\text{Bachelor's}} = \mu_{\text{Graduate}}$$

At least one group has a different number of hours worked per week

$$\mu_i \neq \mu_j$$

b. (1.5 points) Using the information provided, assess whether the following conditions necessary to accurately perform an ANOVA F test are met:
   i. (0.5 point) Are the observations in the study independent?
      i. Yes.
   ii. (0.5 point) Are the sample sizes sufficiently large? (Hint: the *n* row of the table above provides the sample sizes of each group.)
      i. Yes, > 30
   iii. (0.5 point) Is the variation in the groups about equal from one group to the next? (Hint: use the spread of the boxplots and standard deviation values from the table to assess this condition.)
      i. Yes, the SD are similar but different

To assess whether there is a significant difference in the average number of hours worked between one or more of the groups, we need to determine the mean squares between groups (MSTr) and the mean squares within groups (MSE). Each of these values has an associated degrees of freedom.

- The degrees of freedom associated with the MSTr are $df_{Tr} = I - 1 = 5 - 1 = 4$
- The degrees of freedom associated with the MSE are $df_E = N - I = 1172 - 5 = 1167$

c. (1 point) An ANOVA was performed in R. The estimate for the mean squares between groups is MSTr = 501.54 and the resulting *F* statistic is equal to 2.189. Determine the average variation within each group. That is, calculate the MSE.

$$MSE = \frac{MSTr}{F} = \frac{501.54}{2.189}$$

$$MSE = \frac{501.54}{2.189} \approx 229.064$$

d. (2 points) Using the *F* statistic from the previous question (c.) and the two values for the degrees of freedom listed above, calculate the p-value for this test.
0.068, which is greater than 0.05

e. (2 points) Using the p-value calculated in question 5, write a conclusion for this ANOVA F test using a significance level of $\alpha = 0.05$. (Hint: your conclusion should include a statement of evidence in favor of the alternative and a statement as to whether the null hypothesis is rejected or not.)
A p-value of 0.068 suggests that there is some evidence that the average number of hours worked may differ across the groups, but this evidence is not strong enough to be considered significant at the 0.05 level. Since this p-value is higher than the significance level of 0.05, we do not reject the null hypothesis. This means that, statistically, there isn't enough evidence to conclude that the average hours worked differ significantly among the groups.

Gradescope Page Matching (2 points)
When you upload your PDF file to Gradescope, you will need to match each question on this assignment to the correct pages. Video instructions for doing this are available in the Start Here module on Canvas on the page “Submitting Assignments in Gradescope”. Failure to follow these instructions will result in a 2-point deduction on your assignment grade. Match this page to outline item “Gradescope Page Matching”.