```
*******************************
    Exam 2
    Name: Gavin Frias
    Version: 1
*******************************;


*******************************
******** Task 1: DATA *********
*******************************;

/* Question 1: Import Data */
TITLE 'Task1 Q1: Import Data';

%web_drop_table(WORK.IMPORT);


FILENAME REFFILE '/home/u61397358/sasuser.v94/coaster1.csv';

PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV
    OUT=coaster1;
    GETNAMES=YES;
RUN;

PROC CONTENTS DATA=coaster1; RUN;


%web_open_table(WORK.IMPORT);

/* Question 2: Remove the rows that contain missing data (see PDF for column) */
TITLE 'Task1 Q2: Remove Missing Data';

DATA Coaster1_Task1;
    SET Coaster1;
    IF Drop = . THEN DELETE;
RUN;

/* Question 3: Create a new character variable */
TITLE 'Task1 Q3: Create Character Variable';

DATA Coaster1_Task1;
    SET Coaster1;
    LENGTH LengthGroup $6.;
    IF Length<2500 THEN LengthGroup="Short";
    IF Length>=2500 AND Length<4000 THEN LengthGroup="Medium";
    IF Length>=4000 THEN LengthGroup="Long";
RUN;

/* Question 4: Create a new variable called Ratio */
TITLE 'Task1 Q4: Create Ratio';

DATA Coaster1_Task1;
    SET Coaster1;
    Ratio=Height/Drop;
RUN;


/* Question 5: Create a New Dataset called High_Ratio and Print it */
TITLE 'Task1 Q5: Create Dataset High_Ratio';

DATA High_Ratio;
    SET Coaster1_Task1;
    WHERE Ratio>1.15;
    KEEP Track Height Drop Length;
RUN;


************************************************
******** Task 2: INTRODUCTORY ANALYSIS *********
************************************************;

/* Question 6: Compute values of sample mean / median / std dev / IQR
    / # Observations / # Missing */
TITLE 'Task2 Q6: Summary Statistics';

```

```
PROC SORT DATA=Coaster1; by Duration; RUN;
PROC MEANS DATA=Coaster1 MEAN MEDIAN STDDEV QRANGE N NMISS;
by Duration;
RUN;


/* Question 7: Histogram with density kernel */
TITLE 'Task2 Q7: Histogram with Density Kernel';

PROC SGPLOT DATA=Coaster1;
    HISTOGRAM Height;
    DENSITY Height / type=kernel;
RUN;


/* Question 8: Bar Chart */
TITLE 'Task2 Q8: Bar Chart';

PROC SGPLOT DATA=Coaster1;
    VBAR SpeedGroup;
RUN;


/* Question 9: Boxplot */
TITLE 'Task2 Q9: Boxplot';
/* CODE */

PROC SGPLOT DATA=Coaster1;
    HBOX Drop;
RUN;


/*
Are there outliers?
Yes, there is at least one outlier to the far right based on the box plot.
*/


*************************************************
******** Task 3: INFERENCE ********************
*************************************************;

TITLE 'Task3 Q10, Q11: Inference';
    /* CODE */
proc ttest data=Coaster1 h0=0 sides=2 ALPHA=0.017 plots;
var Length;
run;

proc ttest data=Coaster1 ho=-800 sides=u ALPHA=0.017 plots;
var Length;
RUN;

    /* Question 10: Equal Variance Test */
    /*  Hypotheses
            H0: Steel Tracks - Wood Tracks = 0
            H1: Steel Tracks - Wood Tracks != 0
        Test Statistic: 23.31
        P-Value: <0.0001
        Decision: Reject H0
        Conclusion: There is enough evidence to suggest a difference in length between wood and steel track roller coast
    */


    /* Question 11: Mean Testing */
    /*  Hypotheses
            H0: Steel Tracks - Wood Tracks = -800
            H1: Steel Tracks - Wood Tracks < -800
        Test Statistic: 30.23
        P-Value: <0.0001
        Decision: Reject H0
        Conclusion: There is enough evidence to suggest that the mean length of Steel Tracks - Wood Tracks is less than
    */


*************************************************
******** Task 4: REGRESSION ********************
*************************************************;
```

```sas
TITLE 'Task4 Q12: Multiple Linear Regression';
/* CODE */

PROC REG DATA=Coaster1 ALPHA=0.04 ;
    MODEL Duration = Length Type / corrb;
RUN;
/*


Part a - Check model assumptions
    Linearity
        Graph / results looked at: Plot of residuals vs Length and Type.
        Is the linearity condition met or not? Yes.

    Normality
        Graph / results looked at: Plots of residual vs quantile and percent vs residual.
        Is the normality of residuals condition met or not? Yes

    Equal Variance
        Graph / results looked at: Plot of residual vs predicted value
        Is the equal variance of residuals condition met or not? Yes.


Part b - Give the equation of the Multiple Linear Regression line

Duration = B0+B1Length+B2Type
Y = 45.15060 + 0.02386Length + 12.28970Type


Part c - Does the model in total explain variability in Duration?
    Hypotheses
        H0: beta_length = beta_type = 0
        H1: beta_length = beta_type != 0
    Test Statistic: 146.85
    P-Value: <0.0001
    Decision: Reject H0
    Conclusion: There is enough evidence to suggest that at least one variable explains the variability in Duration.


Part d (If needed. If not needed, state why.)

    Testing Individual Variables (Variable 1)
        Hypotheses
            H0: beta_length = 0
            H1: beta_length != 0
        Test Statistic: 16.89
        P-Value: <0.001
        Decision: Reject H0
        Conclusion: There is enough evidence to suggest that Length explains some variability in Duration.


    Testing Individual Variables (Variable 2)
        Hypotheses
            H0: beta_type = 0
            H1: beta_type != 0
        Test Statistic: 1.95
        P-Value: 0.0532
        Decision: Do Not Reject H0
        Conclusion: There is not enough evidence to suggest that Type explains some variability in Duration.


Part e - Value of R^2 and interpretation
    R^2: 0.6835
    Interpretation: We can interpret this as 68.35% of the variability observed in Duration is explained by the model.
*/


*************************************************
******** Task 5: 1-way ANOVA ******************
*************************************************;
TITLE 'Task5 Q13: 1-Way ANOVA';

TITLE2 'Part a: Mean Duration for each Group';
    /* CODE */
PROC MEANS; CLASS SpeedGroup;
RUN;

```

```
    /* Detail any difference by group.
Some differences to note are that the Fast rollercoasters have a higher mean duration
and also have the highest Duration of any rollercoaster.
    */


TITLE2 'Part b: Side by Side Boxpots';
    /* CODE */
PROC SGPLOT DATA=Coaster1;
    HBOX Duration / Category=SpeedGroup;
RUN;


    /* Detail any difference by group.
The boxplot for the Fast variable has several outliers, while the Middle variable had the widest interval.
Another thing to note is that the Small variable boxplot seemed to be the most normal.
    */


TITLE2 'Part c: Run a 1-way ANOVA model';

PROC GLM DATA=Coaster1 ALPHA=0.015;
CLASS SpeedGroup;
MODEL Duration = SpeedGroup;
MEANS SpeedGroup / BON CLDIFF HOVTEST=LEVENE;
OUTPUT OUT = ANOVA13 r = residual;
RUN;

TITLE2 'Part d: Normality Test';
/* Will you test the normality assumption using the overall dataset, or for each group individually?
The overall dataset. */


/* CODE, if needed */
PROC UNIVARIATE NORMAL PLOT DATA=Coaster1 ALPHA=0.015;
VAR Duration;
RUN;


/* Conclusion(s): The data passes the normality check. Shapiro-Wilk = 0.0955 which is greater than 0.05. */


TITLE2 'Part e: Equal Variance Assumption Check';
/* Conclusion: The data passes the equal variance check.*/


TITLE2 'Part f: Is there a significant evidence of an effect?';
/*  Hypotheses
        H0: = 0
        H1: != 0
    Test Statistic: 23.22
    P-Value: <0.0001
    Decision: Reject H0
    Conclusion: There is enough evidence to suggest that Speed Group explains some variability in Duration.
*/


TITLE2 'Part g: Bonerroni or Tukey';
/* Are you providing Bonferroni or Tukey Intervals?
Bonferroni Intervals */


/* Provide confidence intervals for each difference
    (make sure to indicate the difference you are writing a confidence interval for):
Fast - Middle (5.348,55.575)
Fast - Slow (33.885,83.587)
Middle - Fast (-55.575,-5.348)
Middle - Slow (6.655,49.894)
Slow - Fast (-83.587,-33.885)
Slow - Middle (-49.894,-6.655)
*/


/* For each pair, state whether the difference is significant or not
According to my output the difference of each of these confidence intervals are significant.
*/

```

```

TITLE;
```