Course Project

IE 7280 - Statistical Methods in Engineering

Behavioral Analysis in Youth using ANOVA

Monisha Prakash

Professor: Nasser Fard

DATE: 16 August 2019

TOPIC 1- BEHAVIORAL ANALYSIS IN YOUTH USING ANOVA

OBJECTIVE STATEMENT:

The objective of this analysis is to explore the preferences, interests, habits, opinions, and phobias of young people. As per literary research, I found that approximately 10 percent of people in the U.S. experience phobias. In fact, phobias are the most common mental disorder in the U.S. and more women are affected than men. Through our analysis, I will try to identify if women fear certain phenomena significantly more than men. Also, I will be checking if the age group has any significant effect on the phobias. Few of the phobias that I will be testing are darkness, heights, snakes, spiders, stage fright etc.
Therefore, our key objectives are:

1. Whether each of the independent variable would affect the outcome?
2. Are there any interactions between the independent variables?
3. What is the most significant variable contributing to the outcome?

DATA SET DESCRIPTION:

The dataset was chosen from Kaggle. The dataset consists of the survey responses of 1010 students aged 15-30 where they Ire asked about their hobbies, interests, preferences in music and movies, phobias, spending habits etc. Along with their preferences and traits, their demographic information is also captured in the dataset. There are total 150 variables where each variable captures the information of a student's specific trait or preference. 139 variables are numerical and 11 are categorical. Information about 10 phobias is captured in the dataset, each as a separate variable. Therefore, I have 10 dependent variables. The responses are ordinal in nature, ranging from 1-5 which signifies not afraid of - very afraid of. Age is taken as one of the predictors and it has values ranging from 15-30. I binned the discrete ages into age groups, 15-20, 20-25 and 25-30.

Source: https://www.kaggle.com/miroslavsabo/young-people-survey

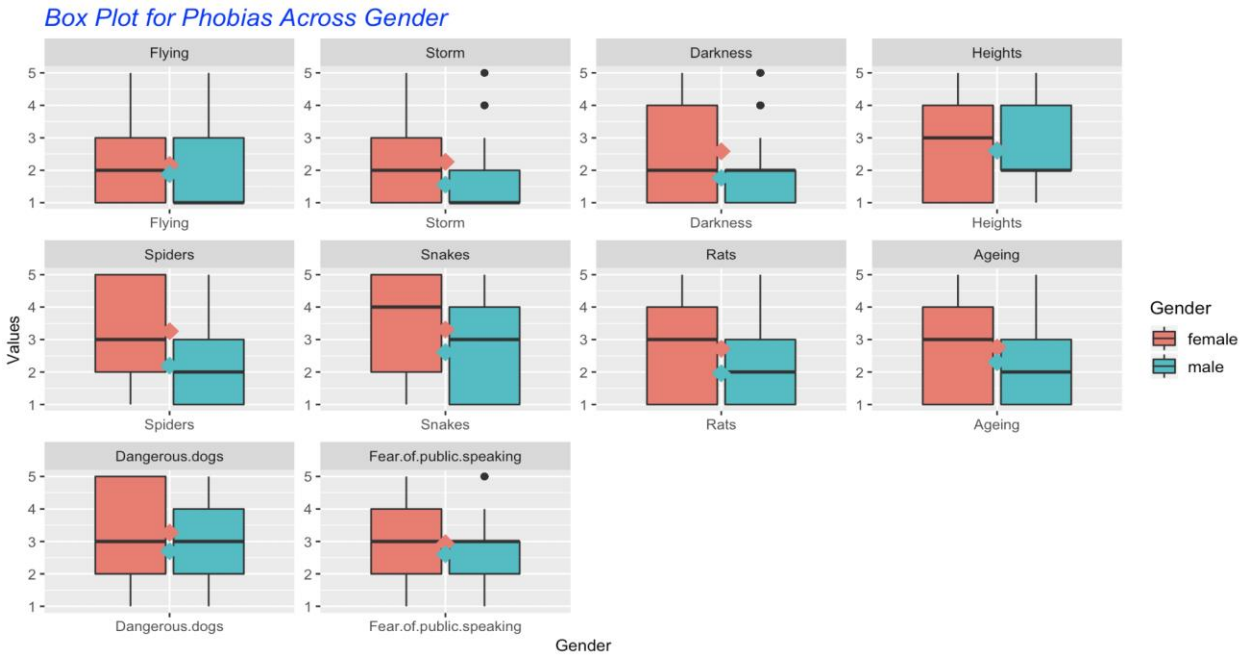
OUTLINE OF STATISTICAL ANALYSIS PROCEDURE:

This study applies 2 – way ANOVA for analyzing the effect of Gender and Age on Phobias. Test for Hypothesis was performed to check if there's any significant difference between the variables and to check if there is presence of a significant dependence of phobias on age or gender. Also, Tukey's test is performed to indicate the significant variable.

DATA EXPLORATION:

	vars	n	mean	sd	min	max	range	se	missing_values
Parents..advice	131	1008	3.265873	0.8657364	1	5	4	0.02726813	2
Questionnaires.or.polls	132	1006	2.748509	1.1015018	1	5	4	0.03472852	4
Internet.usage*	133	1010	1.392079	0.7086725	1	4	3	0.02229897	0
Finances	134	1007	3.023833	1.1443647	1	5	4	0.03606199	3
Shopping.centres	135	1008	3.234127	1.3230618	1	5	4	0.04167253	2
Branded.clothing	136	1008	3.050595	1.3063209	1	5	4	0.04114524	2
Entertainment.spending	137	1007	3.201589	1.1889474	1	5	4	0.03746691	3
Spending.on.looks	138	1007	3.106256	1.2053685	1	5	4	0.03798438	3
Spending.on.gadgets	139	1010	2.870297	1.2849703	1	5	4	0.04043267	0
Spending.on.healthy.eating	140	1008	3.557540	1.0937498	1	5	4	0.03444988	2
Age	141	1003	20.433699	2.8288401	15	30	15	0.08932190	7
Height	142	990	173.514141	10.0245050	62	203	141	0.31859968	20
Weight	143	990	66.405051	13.8395608	41	165	124	0.43985012	20
Number.of.siblings	144	1004	1.297809	1.0133482	0	10	10	0.03198099	6
Gender*	145	1010	2.400990	0.5023227	1	3	2	0.01580600	0
Left...right.handed*	146	1010	2.894059	0.3174233	1	3	2	0.00998799	0
Education*	147	1010	5.650495	2.0171772	1	7	6	0.06347217	0
Only.child*	148	1010	2.249505	0.4374953	1	3	2	0.01376616	0
Village...town*	149	1010	2.292079	0.4635758	1	3	2	0.01458680	0
House...block.of.flats*	150	1010	2.402970	0.4987508	1	3	2	0.01569361	0

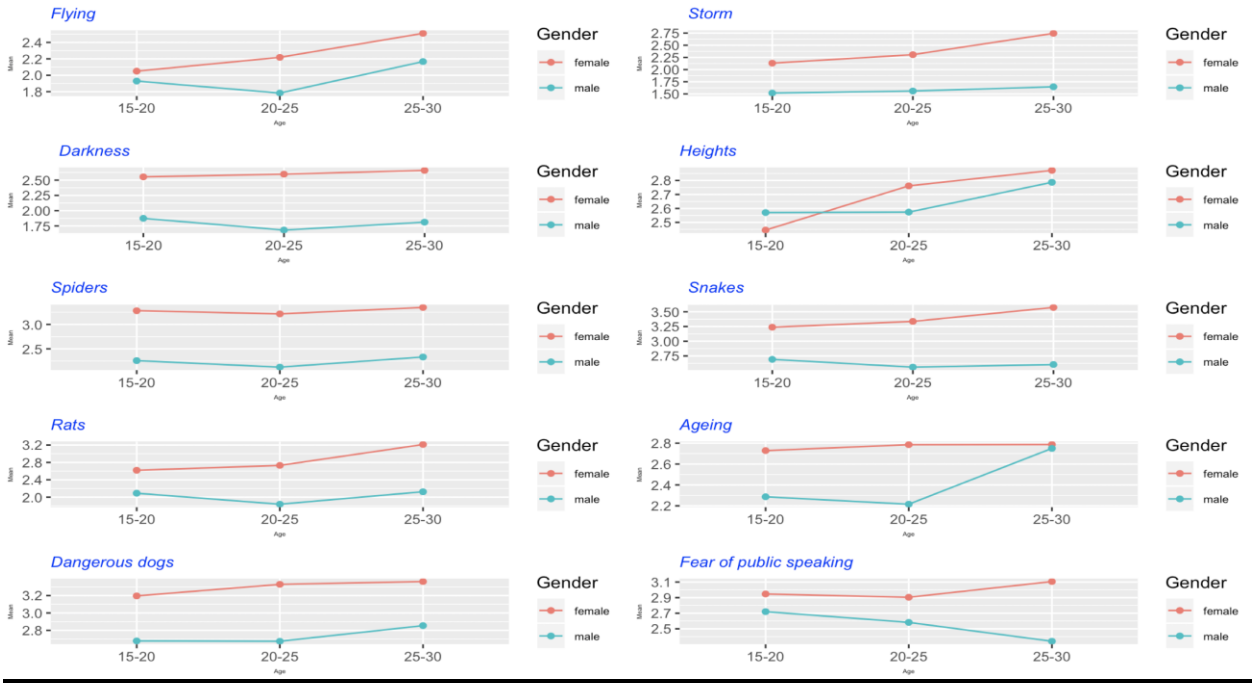
From the summary statistics I can see that there 1010 data points. Since I will be analyzing across Gender and Age Group, I do not want missing values across these variables. As I can see in the last column that Gender does not have any missing values. However, age information is missing for total 7 records. These records Ire deleted before proceeding with our analysis.



I plotted box plots to analyze the distribution of male and female for different phobias and to also see the difference in their means. I can see that the mean for females is higher in comparison to males for all phobias except for Heights. This analysis helps us in stating our hypothesis that phobias are gender dependent. Also, I can see that the data is more skwed for males in comparison to females.

I plotted box plots to analyze the distribution of different age groups for different phobias. I can see that the data is not much skewed for most of the phobias except for Spiders, Snakes, Rats and Ageing. Also, there is not much difference between the means of the 3 age groups. Fear of flying, storm, heights and ageing increases with age. However, hear of darkness, spiders, stage fright decreases with age. There seem to be a minute difference and I will be checking the significance of age group of a person on the phobias.

Interaction betIen Age and Gender



I do not see much interaction between age and gender across different phobias. I see interaction betIen age and gender for the fear of heights, ageing and flying up to a little extent. Later, I will be testing the significance of interaction between age and gender using ANOVA.

From the interaction plots, I see that the phobias for male and female are not dependent on their age as the lines for most of the phobias for both male and female are parallel. However, phobias in female are more prevalent in comparison to males

ANOVA OF MODEL

Flying:

```
             Df Sum Sq Mean Sq F value   Pr(>F)
Age_Group     2    8.9   4.450   3.128 0.044227 *
Gender        1   21.4  21.352  15.011 0.000114 ***
Residuals   991 1409.6   1.422
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Darkness:

```
             Df Sum Sq Mean Sq F value Pr(>F)
Age_Group     2    4.5    2.27   1.605  0.201
Gender        1  157.1  157.12 111.047 <2e-16 ***
Residuals   992 1403.6    1.41
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Heights:

```
             Df Sum Sq Mean Sq F value Pr(>F)
Age_Group     2   13.0   6.503   3.899 0.0206 *
Gender        1    0.7   0.657   0.394 0.5304
Residuals   991 1653.0   1.668
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Spiders:

```
              Df Sum Sq Mean Sq F value Pr(>F)
Age_Group     2   10.5    5.24   2.468 0.0852 .
Gender        1  263.2  263.19 123.883 <2e-16 ***
Residuals   989 2101.1    2.12
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Storm:

```
              Df Sum Sq Mean Sq F value Pr(>F)
Age_Group     2    5.5    2.73   2.228  0.108
Gender        1  126.2  126.17 102.912 <2e-16 ***
Residuals   993 1217.5    1.23
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Snakes:

```
              Df Sum Sq Mean Sq F value   Pr(>F)
Age_Group     2    1.5    0.76   0.356    0.701
Gender        1  117.7  117.75  54.970 2.62e-13 ***
Residuals   994 2129.2    2.14
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Rats:

```
              Df Sum Sq Mean Sq F value Pr(>F)
Age_Group     2    9.9    4.94   2.726  0.066 .
Gender        1  138.9  138.88  76.660 <2e-16 ***
Residuals   991 1795.4    1.81
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ageing:

```
              Df Sum Sq Mean Sq F value   Pr(>F)
Age_Group     2    4.7    2.35   1.262    0.284
Gender        1   51.4   51.38  27.610 1.82e-07 ***
Residuals   993 1848.0    1.86
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Dangerous Dogs:

```
              Df Sum Sq Mean Sq F value   Pr(>F)
Age_Group      2    0.6    0.28   0.155    0.857
Gender         1   81.4   81.37  45.419 2.69e-11 ***
Residuals    993 1778.9    1.79
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fear of Public Speaking:

```
              Df Sum Sq Mean Sq F value   Pr(>F)
Age_Group      2    3.5   1.746   1.208    0.299
Gender         1   25.4  25.440  17.602 2.97e-05 ***
Residuals    993 1435.1   1.445
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Insights

Looking at the summary table for various phobias across various age and gender, I can say that out of the 10 phobias, only flying and heights are dependent on age. HoIver, all phobias except for heights are gender dependent. Also, I can deduce that gender is significant at different confidence intervals.

TEST OF HYPOTHESIS FOR THESE VARIABLES

To avoid redundancy, I will be showing Hypothesis Test only for few phobias. I are taking 95% Confidence Interval for the Hypothesis test

1. **Gender:**
   a. **Flying:**
      $H0$: $u1 = u2$
      $H1$: $u1 \neq u2$
      P_value = 0.000114< 0.05
      $\therefore$ $Reject\ H0$. There is enough evidence to prove that fear of flying is dependent on gender
   b. **Heights:**
      $H0$: $u1 = u2$
      $H1$: $u1 \neq u2$
      P_value = 0.5304 > 0.05
      $\therefore$ Fail to r$eject\ H0$. I do not have enough evidence to prove that fear of heights is dependent on gender
   c. **Rats:**
      $H0$: $u1 = u2$
      $H1$: $u1 \neq u2$
      P_value = `2e-16`< 0.05
      $\therefore$ $Reject\ H0$. There is enough evidence to prove that fear of rats is dependent on gender
   d. **Ageing:**

$H0$: $u1 = u2$
$H1$: $u1 \neq u2$
P_value = 1.82e-07< 0.05
∴ $Reject\ H0$. There is enough evidence to prove that fear of ageing is dependent on gender

e. <u>Fear of Public Speaking:</u>

$H0$: $u1 = u2$
$H1$: $u1 \neq u2$
P_value = 2.97e-05< 0.05
∴ $Reject\ H0$. There is enough evidence to prove that fear of public speaking is dependent on gender

2. <u>Age Group:</u>

a. <u>Flying:</u>

$H0$: $u1 = u2 = u3$
$H1$: $At\ least\ two\ ui's\ are\ not\ the\ same.\ i = 1\ to\ 3$
P_value = 0.044227< 0.05
∴ $Reject\ H0$. There is enough evidence to prove that fear of flying is dependent on age of a person

b. <u>Heights:</u>

$H0$: $u1 = u2 = u3$
$H1$: $At\ least\ two\ ui's\ are\ not\ the\ same.\ i = 1\ to\ 3$
P_value = 0.0206< 0.05
∴ $Reject\ H0$. There is enough evidence to prove that fear of heights is dependent on age of a person

c. <u>Rats:</u>

$H0$: $u1 = u2 = u3$
$H1$: $At\ least\ two\ ui's\ are\ not\ the\ same.\ i = 1\ to\ 3$
P_value = 0.066> 0.05
∴ Fail to r$eject\ H0$. There isn't enough evidence to prove that fear of rats is dependent on age of a person

d. <u>Ageing:</u>

$H0$: $u1 = u2 = u3$
$H1$: $At\ least\ two\ ui's\ are\ not\ the\ same.\ i = 1\ to\ 3$
P_value = 0.284>0.05
∴ Fail to r$eject\ H0$. There isn't enough evidence to prove that fear of ageing is dependent on age of a person

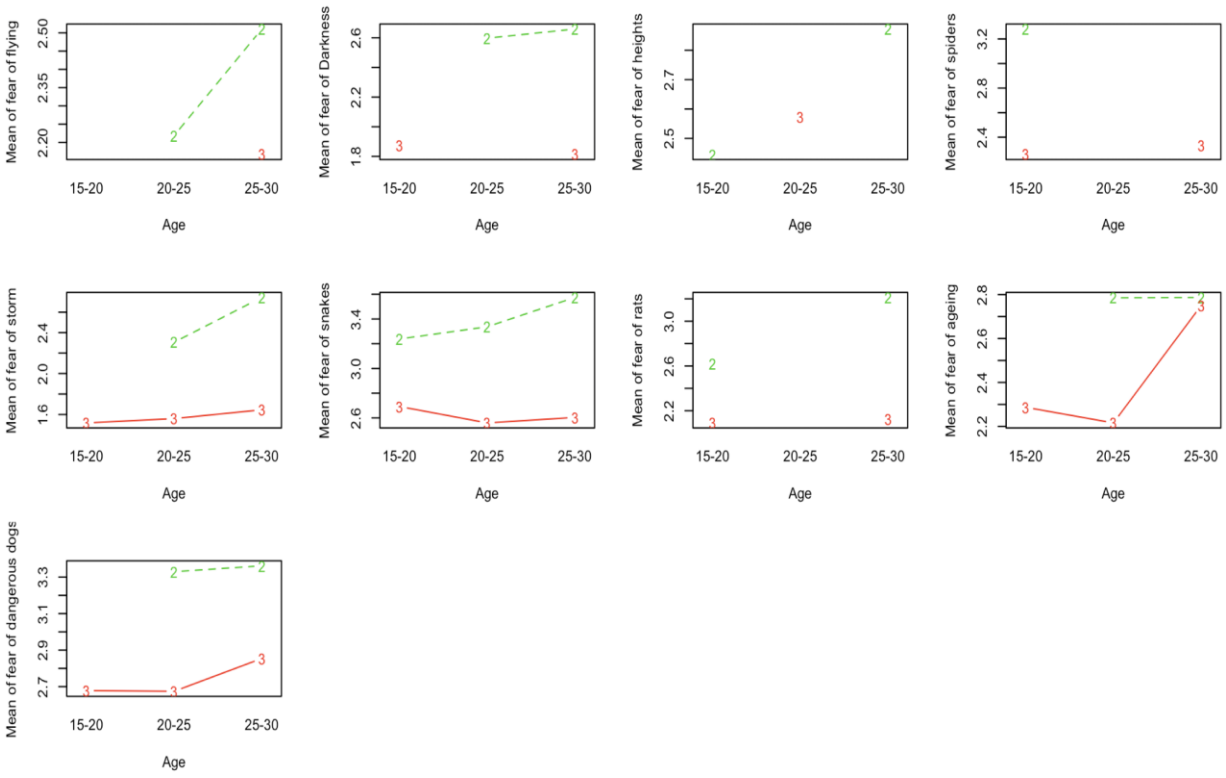
e. <u>Fear of Public Speaking:</u>

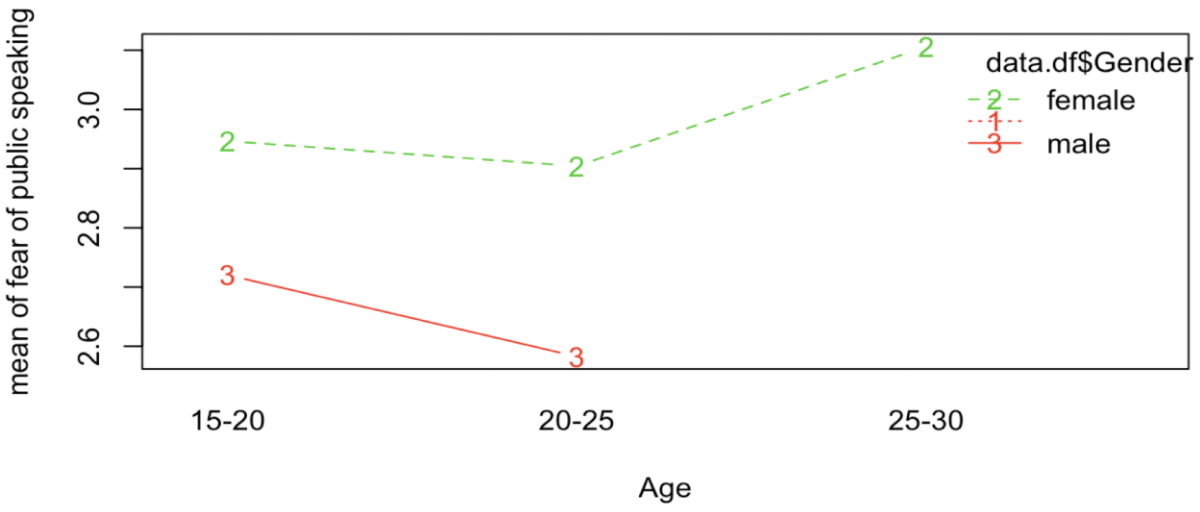
$H0$: $u1 = u2 = u3$
$H1$: $At\ least\ two\ ui's\ are\ not\ the\ same.\ i = 1\ to\ 3$
P_value = 0.299> 0.05
∴ Fail to r$eject\ H0$. There isn't enough evidence to prove that fear of public speaking is dependent on age of a person.

RESULTS OF 2-WAY ANOVA

Furthermore, to determine the effect of individual variables when in contact with other significant contributors 2-way ANOVA is used.

Interaction Plots





From the interaction plots, I see that the phobias for male and female are not dependent on their age as the lines for most of the phobias for both male and female are parallel. However, phobias in female are more prevalent in comparison to males.

From the previous results, I saw that it was just the fear of flying where both gender and age group Ire significant predictors. However, I checked the interaction between age and gender for all phobias and I did not find any significant interaction at 0.05 confidence level.

```
                   Df Sum Sq Mean Sq F value   Pr(>F)
Age_Group           2    8.9   4.450   3.134 0.043993 *
Gender              1   21.4  21.352  15.036 0.000112 ***
Age_Group:Gender    2    5.2   2.622   1.846 0.158390
Residuals         989 1404.4   1.420
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

I can see that individually both the variables are significant. However, their interaction is not significant as the p value is greater than 0.05.

TUKEY’S TEST

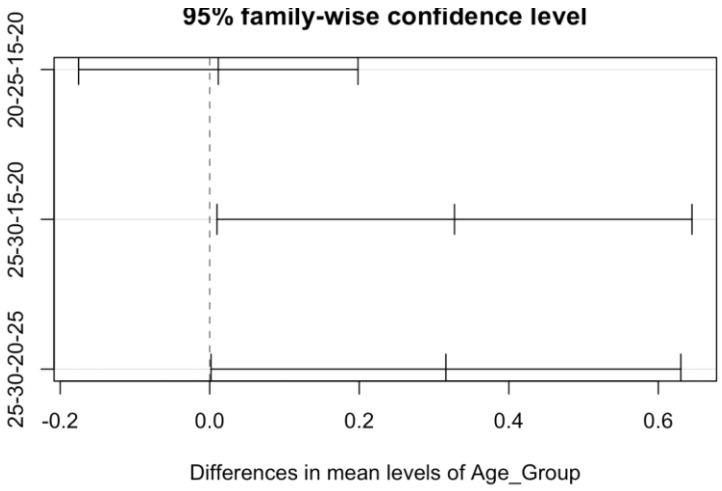
The Tukey Test (or Tukey procedure), also called Tukey’s Honest Significant Difference test, is a post-hoc test based on the studentized range distribution. An ANOVA test can tell you if your results are significant overall, but it won’t tell you exactly where those differences lie. After you have run an ANOVA and found significant results, then you can run Tukey’s HSD to find out which specific groups means (compared with each other) are different. The test compares all possible pairs of means.

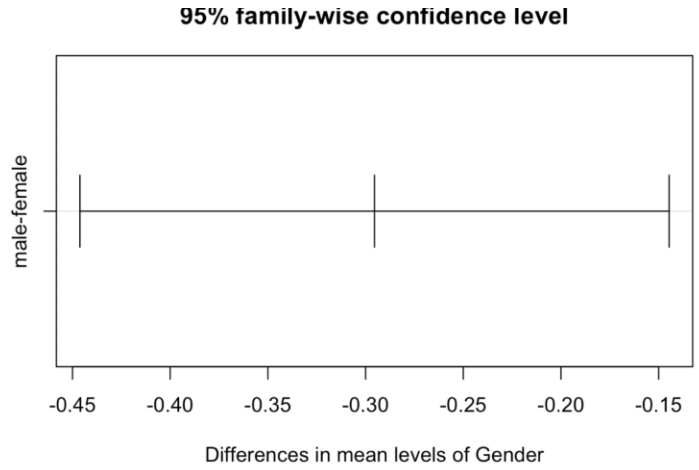
Assumptions for the test
- Observations are independent within and among groups.
- The groups for each mean in the test are normally distributed.
- There is equal within-group variance across the groups associated with each mean in the test (homogeneity of variance)

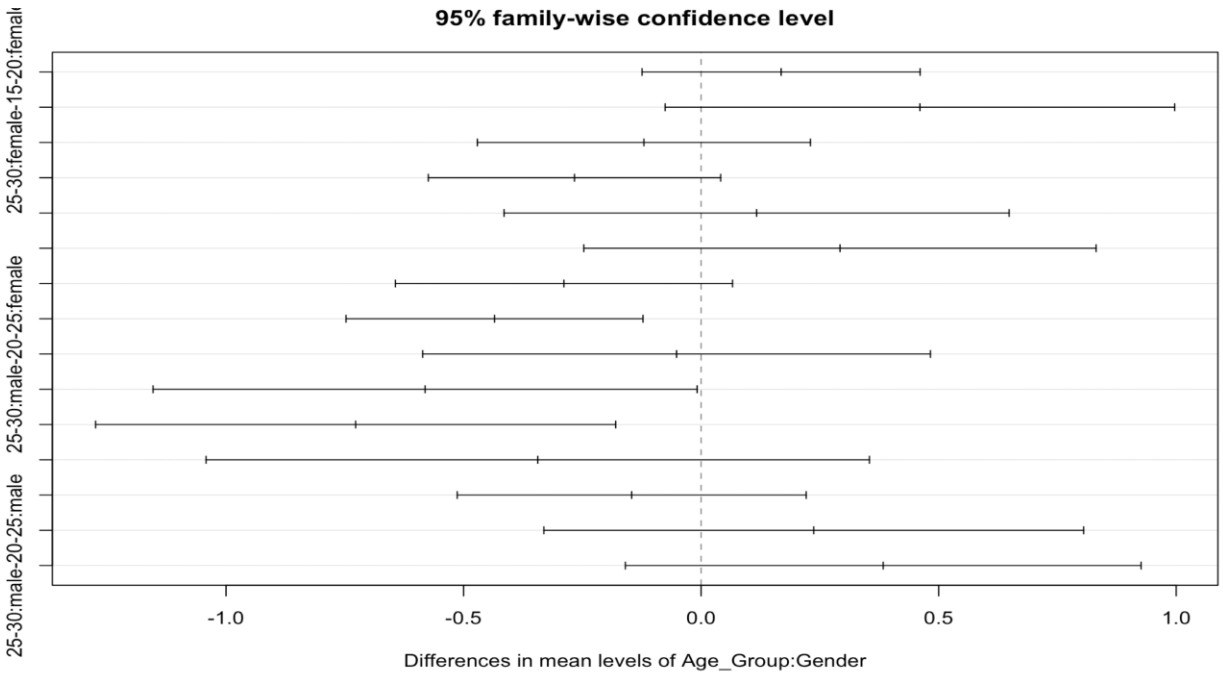
I performed Tukey’s test for the fear of Flying:

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

Fit: aov(formula = Flying ~ Age_Group + Gender + Age_Group:Gender, data
= data.df[data.df$Gender != "", ])

$Age_Group
                   diff          lwr       upr     p adj
20-25-15-20 0.01144183 -0.175390831 0.1982745 0.9886730
25-30-15-20 0.32736343  0.009727986 0.6449989 0.0416073
25-30-20-25 0.31592160  0.001723704 0.6301195 0.0483998
```

```
$Gender
                 diff        lwr        upr     p adj
male-female -0.2954206 -0.4462026 -0.1446385 0.0001284

$`Age_Group:Gender`
                                    diff         lwr          upr     p adj
20-25:female-15-20:female   0.16839080 -0.1243592  0.461140772 0.5704995
25-30:female-15-20:female   0.46063830 -0.0757057  0.996982295 0.1396501
15-20:male-15-20:female    -0.12042254 -0.4709568  0.230111773 0.9240338
20-25:male-15-20:female    -0.26658986 -0.5743177  0.041137989 0.1329532
25-30:male-15-20:female     0.11666667 -0.4148719  0.648205232 0.9890540
25-30:female-20-25:female   0.29224749 -0.2468951  0.831390117 0.6332844
15-20:male-20-25:female    -0.28881334 -0.6436150  0.065988277 0.1852212
20-25:male-20-25:female    -0.43498067 -0.7475608 -0.122400570 0.0010691
25-30:male-20-25:female    -0.05172414 -0.5860865  0.482638223 0.9997828
15-20:male-25-30:female    -0.58106083 -1.1536397 -0.008481994 0.0443671
20-25:male-25-30:female    -0.72722816 -1.2746482 -0.179808160 0.0021781
25-30:male-25-30:female    -0.34397163 -1.0421876  0.354244290 0.7230525
20-25:male-15-20:male      -0.14616733 -0.5134248  0.221090177 0.8661419
25-30:male-15-20:male       0.23708920 -0.3309908  0.805169205 0.8409485
25-30:male-20-25:male       0.38325653 -0.1594561  0.925969178 0.3336097
```





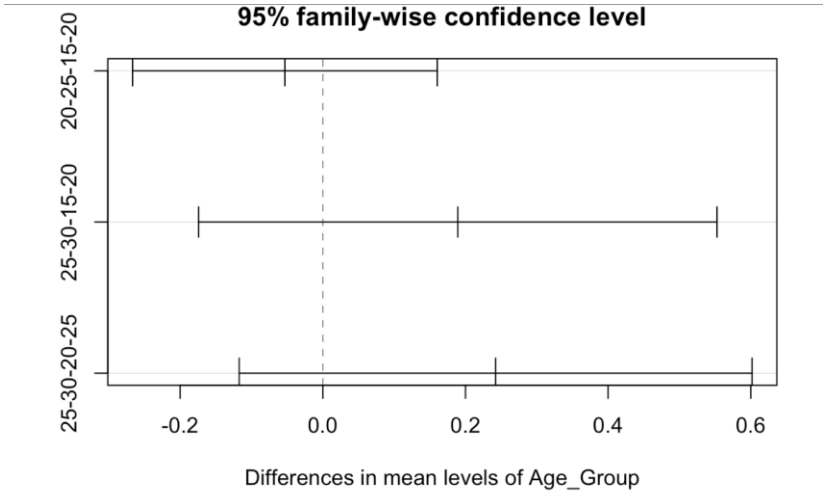


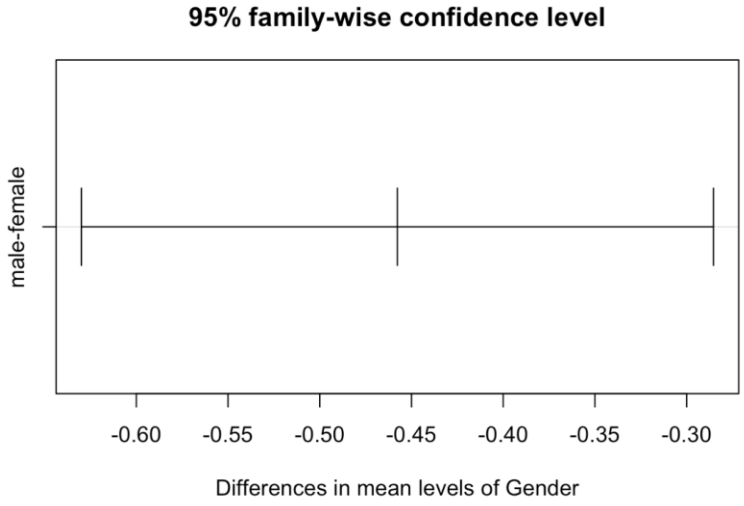
I can see that there is a significant difference in mean between males and females. There is a difference of 0.3 between their means. This proves that females fear flying more than males. Also, there is a significant difference in the fear of flying for people in the age groups 25-30 and 15-30 and the difference is of 0.33 and between age groups 20-25 and 25-30. This shows that fear of flying increases with age. The interaction is not significant. However, I see that there is a significant difference in means between male and female for the age group 20-25 implying that for the age group 20-25, females fear flying more than men. Also, difference between 20-25 male and 25-30 female is significant which tells us that since the person is older and is a female, the tendency of having flying phobia is higher.

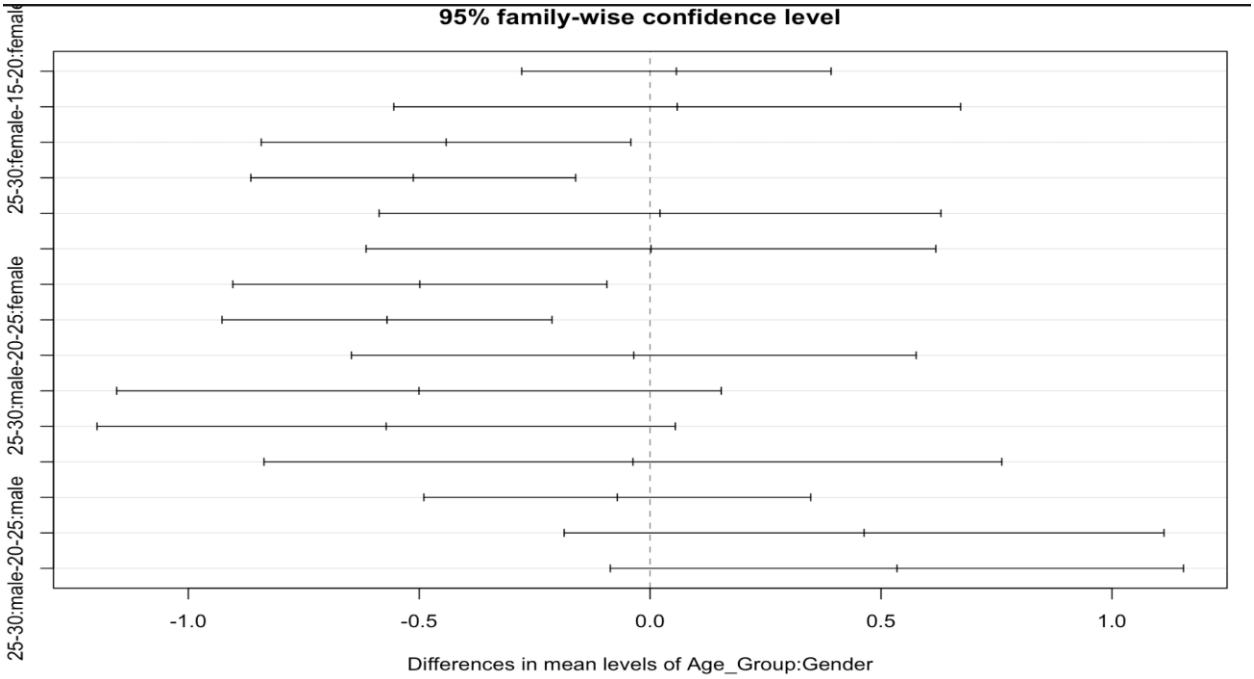
I did a similar test for <u>Ageing</u>:

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

```
Fit: aov(formula = Ageing ~ Age_Group + Gender + Age_Group:Gender, data =
data.df[data.df$Gender != "", ])

$Age_Group
                    diff        lwr       upr     p adj
20-25-15-20 -0.05310018 -0.2666355 0.1604352 0.8288929
25-30-15-20  0.18922484 -0.1741353 0.5525850 0.4401960
25-30-20-25  0.24232502 -0.1171186 0.6017687 0.2537007

$Gender
                  diff        lwr        upr p adj
male-female -0.4576743 -0.6299501 -0.2853985 2e-07

$`Age_Group:Gender`
                                      diff          lwr          upr     p adj
20-25:female-15-20:female       0.05686918 -0.27809562  0.39183399 0.9966965
25-30:female-15-20:female       0.05866261 -0.55502273  0.67234795 0.9997957
15-20:male-15-20:female        -0.44185814 -0.84200828 -0.04170801 0.0206084
20-25:male-15-20:female        -0.51297510 -0.86462235 -0.16132785 0.0004832
25-30:male-15-20:female         0.02142857 -0.58675839  0.62961553 0.9999986
25-30:female-20-25:female       0.00179343 -0.61509410  0.61868096 1.0000000
15-20:male-20-25:female        -0.49872733 -0.90377135 -0.09368330 0.0060942
20-25:male-20-25:female        -0.56984428 -0.92705055 -0.21263801 0.0000860
25-30:male-20-25:female        -0.03544061 -0.64685856  0.57597734 0.9999828
15-20:male-25-30:female        -0.50052076 -1.15509614  0.15405463 0.2464298
20-25:male-25-30:female        -0.57163771 -1.19774041  0.05446499 0.0964844
25-30:male-25-30:female        -0.03723404 -0.83613340  0.76166531 0.9999942
20-25:male-15-20:male          -0.07111696 -0.49006209  0.34782818 0.9966987
25-30:male-15-20:male           0.46328671 -0.18613658  1.11271001 0.3220773
25-30:male-20-25:male           0.53440367 -0.08631066  1.15511800 0.1377422
```




95% family-wise confidence level
male-female
-0.60
-0.55
-0.50
-0.45
-0.40
-0.35
-0.30
Differences in mean levels of Gender



Analyzing the results of Tukey's test for the fear of ageing, I can see that there are no 2 age groups whose mean difference is significant. However, I see a significant difference of 0.45 between males and females i.e. the fearing in male is ~0.5 lesser in males in comparison to females (p= `2e-07`). Age group doesn't have any significant effect on phobia of ageing. Also, interaction between the 2 independent variables in not significant.

CONCLUSION

A two-way ANOVA was performed to test the effect of Gender and Age Group independently on various phobias. Also, the effect of their interaction was also analyzed. From the above observations, looking at Tukey's test and ANOVA values I can conclude that Gender and phobias are not independent at 0.05 level of significance. Apart from heights, females have the tendency to fear more in comparison to males. Age Group is a significant predictor only for the fear of heights and flying. It is not a significant predictor for other phobias. The significance of their interaction was tested, and I see that their interaction is not relevant.