**Name**: Vishwa Koppisetti

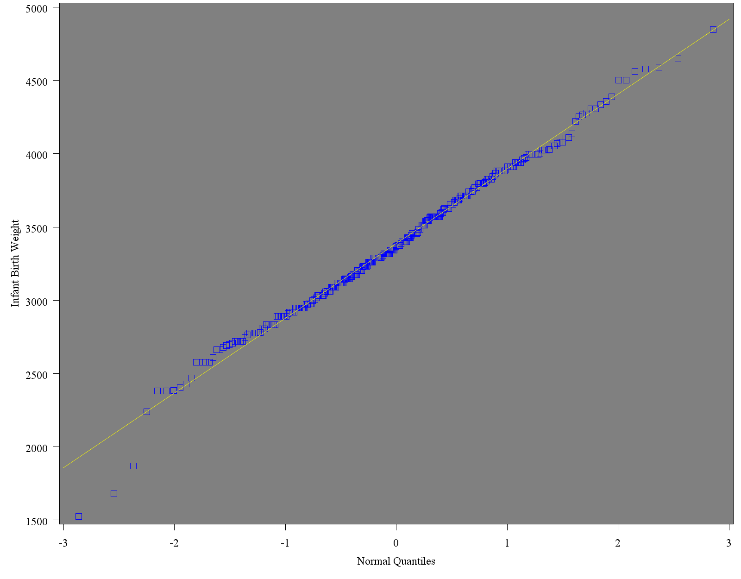
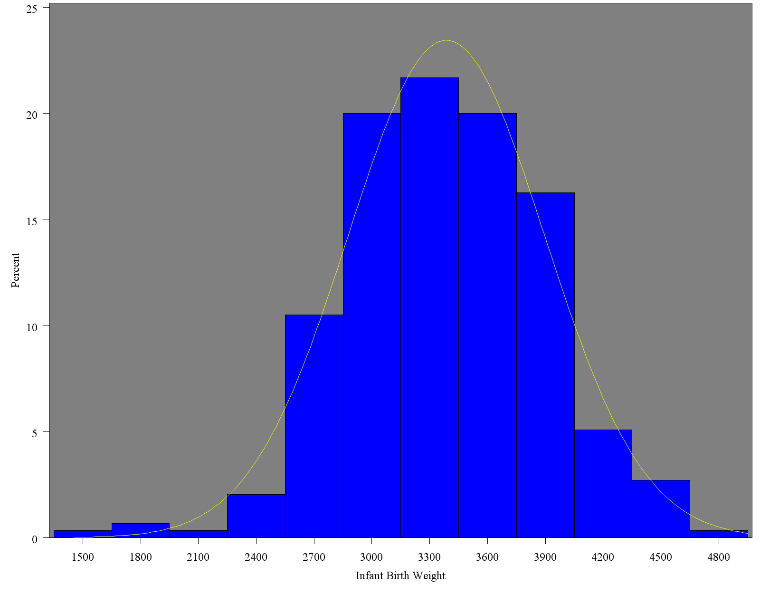
**ID**: qte947

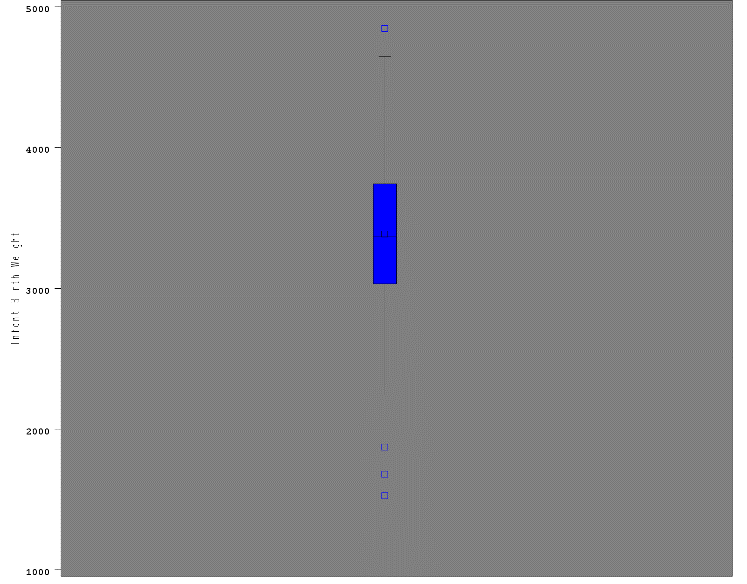
# midterm

## Exercise 1:

1. **Qualitative/visual analysis**: The histogram shows bell-shaped curve with normal-like distribution, slightly left skewed or negatively skewed distribution. The Q-Q plot shows fairly normal distribution since we observe a straight line between the first and third quantiles. The box plot shows fairly normal distribution with the mean nearly the same as median but mean slightly greater. All three plots show the presence of few outliers.

**General features of distribution**: The mean and median are nearly the same with mean being slightly greater than median. This implies slightly left skewed distribution which is supported by the skewness value (-0.106) and also seen in histogram. The kurtosis is positive (0.525) and close to 0 hence the histogram shows mostly curved peak and not too heavy tails but fairly normal distribution. The variability or spread of data is uniform as seen in the histogram and box plot.





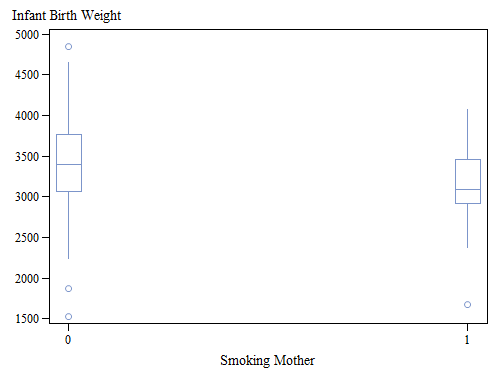
**Quantitative analysis**: The Shapiro-Wilk test yields a p-value 0.1153 which is not significant at p < 0.05 hence we retain null hypothesis that **normality assumption is met** for infant birth weight (**Weight**)

| **Tests for Normality** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Shapiro-Wilk** | **W** | 0.992057 | **Pr < W** | 0.1153 |
| **Kolmogorov-Smirnov** | **D** | 0.032207 | **Pr > D** | >0.1500 |
| **Cramer-von Mises** | **W-Sq** | 0.04358 | **Pr > W-Sq** | >0.2500 |
| **Anderson-Darling** | **A-Sq** | 0.378672 | **Pr > A-Sq** | >0.2500 |

1. **Generate boxplot of Weight by MomSmoke and informally compare infant birth weights between levels.**

MomSmoke has 2 levels : 0 is non-smoking , 1 is smoking mom

**Infant Birth weight is more for Non-smoking moms than for smoking moms**. The data seems to be fairly normally distributed widely spread for non-smoking moms whereas for smoking moms there seems median is closer to the lower quartile which suggests left skewed distribution.



1. **For each level in MomSmoke, quantitatively test for normality of Weight and make a conclusion.**

MomSmoke has 2 levels: 0 is non-smoking, 1 is smoking mom.

For non-smoking mom, the weights are normally distributed since the Shapiro-Wilk (p-value 0.3549) test is not significant at p < 0.05 We can retain null hypothesis that assumption of normality is reasonable.

| **Tests for Normality** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Shapiro-Wilk** | **W** | 0.99362 | **Pr < W** | 0.3549 |
| **Kolmogorov-Smirnov** | **D** | 0.028159 | **Pr > D** | >0.1500 |
| **Cramer-von Mises** | **W-Sq** | 0.027532 | **Pr > W-Sq** | >0.2500 |
| **Anderson-Darling** | **A-Sq** | 0.251243 | **Pr > A-Sq** | >0.2500 |

For smoking mom, the weights are normally distributed since the Shapiro-Wilk (p-value 0.2000) test is not significant at p < 0.05 . We can retain null hypothesis that assumption of normality is reasonable.

| **Tests for Normality** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Shapiro-Wilk** | **W** | 0.962992 | **Pr < W** | 0.2000 |
| **Kolmogorov-Smirnov** | **D** | 0.098235 | **Pr > D** | >0.1500 |
| **Cramer-von Mises** | **W-Sq** | 0.074047 | **Pr > W-Sq** | 0.2444 |
| **Anderson-Darling** | **A-Sq** | 0.464695 | **Pr > A-Sq** | 0.2458 |

## Exercise 2:

We can use parametric test since both levels of MomSmoke satisfy Normality assumption as seen above in exercise 1. We can use two sample t-test since there are only 2 groups/levels in the categorical variable (MomSmoke).

**Homogeneity of variance assumption is met** from Folded F test. p-value 0.8009 is insignificant at p < 0.05, we retain null hypothesis that variance are equal. Hence, we take p-value 0.0023 from pooled variance test and decide that since p-value is significant at p < 0.05, we reject null hypothesis and **retain alternative hypothesis that the infants from smoking mom have significantly different weights than those from non-smoking mom.**

| **Equality of Variances** | | | | |
| --- | --- | --- | --- | --- |
| **Method** | **Num DF** | **Den DF** | **F Value** | **Pr > F** |
| **Folded F** | 253 | 40 | 1.08 | 0.8009 |

| **Method** | **Variances** | **DF** | **t Value** | **Pr > |t|** |
| --- | --- | --- | --- | --- |
| **Pooled** | Equal | 293 | 3.07 | 0.0023 |
| **Satterthwaite** | Unequal | 54.877 | 3.16 | 0.0026 |

From the below table it is apparent that **non-smoking moms have infants with more birth weight than smoking moms.**

| **MomSmoke** | **N** | **Mean** | **Std Dev** | **Std Err** | **Minimum** | **Maximum** |
| --- | --- | --- | --- | --- | --- | --- |
| **0** | 254 | 3422.7 | 505.6 | 31.7233 | 1526.0 | 4848.0 |
| **1** | 41 | 3162.7 | 486.8 | 76.0296 | 1680.0 | 4070.0 |
| **Diff (1-2)** |  | 260.0 | 503.1 | 84.6697 |  |  |

## Exercise 3:

1. **Check homogeneity of variance assumption. Does it hold and okay to perform ANOVA?**

**Homogeneity of variance assumption**: Levene’s test shows a p-value 0.7271 which is not significant at p < 0.05 Hence we retain the null hypothesis that homogeneity of variance assumption is met. We also checked that normality assumption is met hence we can perform ANOVA.

| **Levene's Test for Homogeneity of Weight Variance ANOVA of Squared Deviations from Group Means** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **MomSmoke** | 1 | 1.932E10 | 1.932E10 | 0.12 | 0.7271 |
| **Error** | 293 | 4.639E13 | 1.583E11 |  |  |

1. **Make a conclusion on the effect of MomSmoke. Compare your result with conclusion of Exercise 2.**

From the model we can observe the p-value to be 0.0023 which is significant at p < 0.05 hence we reject null hypothesis and conclude that MomSmoke has significant effect on infant birth weight. **This is in agreement with the conclusion of Exercise 2.** Also the model explains only 3.1% variation in the infant birth weight.

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Model** | 1 | 2386708.35 | 2386708.35 | 9.43 | 0.0023 |
| **Error** | 293 | 74151291.20 | 253076.08 |  |  |
| **Corrected Total** | 294 | 76537999.55 |  |  |  |

| **R-Square** | **Coeff Var** | **Root MSE** | **Weight Mean** |
| --- | --- | --- | --- |
| 0.031183 | 14.85468 | 503.0667 | 3386.586 |

Since there are only 2 levels we can conclude that non-smoking moms have infants with more birth weight than smoking moms. **This agrees with the conclusion of Exercise 2.**

| **Level of MomSmoke** | **N** | **Weight** | |
| --- | --- | --- | --- |
| **Mean** | **Std Dev** |
| **0** | 254 | 3422.72441 | 505.586509 |
| **1** | 41 | 3162.70732 | 486.826676 |

## Exercise 4:

1. We run a 5-way ANOVA for 5 explanatory variables and one dependent variable. For backward selection we perform the following steps:

* We look for least significant term (a factor with highest p-value that is > 0.05) based on Type 3 SS. We **remove Ed** effect which is least significant at p < 0.05 with p-value 0.8626
* We run 4-way ANOVA with remaining 4 factors (black, married, boy, MomSmoke). We **remove married** effect as it is the least significant term of the 4 effects with p-value 0.6158 not significant at p < 0.05
* We run 3-way ANOVA with remaining 3 effects (black, boy, MomSmoke). We **remove boy** effect as it is the least significant term of the 3 effects with p-value 0.3888 not significant at p < 0.05
* We finally keep the 2 significant terms Black and MomSmoke with p-values 0.0001 and 0.0014 respectively. We run 2-way ANOVA with main effects Black, MomSmoke and their interaction and get the below final 2-way ANOVA model.

| * **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Model** | 3 | 6049211.70 | 2016403.90 | 8.32 | <.0001 |
| **Error** | 291 | 70488787.85 | 242229.51 |  |  |
| **Corrected Total** | 294 | 76537999.55 |  |  |  |

| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Black** | 1 | 1467232.670 | 1467232.670 | 6.06 | 0.0144 |
| **MomSmoke** | 1 | 1161829.335 | 1161829.335 | 4.80 | 0.0293 |
| **Black\*MomSmoke** | 1 | 5461.337 | 5461.337 | 0.02 | 0.8807 |

The interaction term (p-value 0.8807) is not significant at p < 0.05. There is no significant interaction between Black and MomSmoke. The variation is Black factor is same for both levels of MomSmoke and vice versa.

1. The final model is a 2-way ANOVA with main effects Black and MomSmoke. The interaction term is not significant, so it is removed. The model is significant at p < 0.05 with a p-value < .0001.

Only 7.8% of variation in the infant birth weight is explained by the model.

| 1. **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Model** | 2 | 6043750.36 | 3021875.18 | 12.52 | <.0001 |
| **Error** | 292 | 70494249.18 | 241418.66 |  |  |
| **Corrected Total** | 294 | 76537999.55 |  |  |  |

| **R-Square** | **Coeff Var** | **Root MSE** | **Weight Mean** |
| --- | --- | --- | --- |
| 0.078964 | 14.50852 | 491.3437 | 3386.586 |

**ANOVA Assumptions through diagnostic plots**

The Residuals show normal distribution in histogram and straight line in Q-Q plot, so assumption of normality is reasonable. The r-student shows equal assumption of variance as the data points fall between values -2 and 2



1. **State conclusions about significant differences in Weight across groups. For each significant variable, state specifically which category has a larger or smaller mean value of Weight.**

Black predictor variable has a significant effect on infant birth weight. Infants from White people have more birth weight than infants from black people. **White people have larger mean value of weight than black people**. The mean weight difference is significant as 95% confidence interval doesn’t include 0.

| **Black** | **Weight LSMEAN** | **95% Confidence Limits** | |
| --- | --- | --- | --- |
| **0** | 3339.939710 | 3255.129873 | 3424.749546 |
| **1** | 3040.706892 | 2889.504691 | 3191.909093 |

| **Least Squares Means for Effect Black** | | | | |
| --- | --- | --- | --- | --- |
| **i** | **j** | **Difference Between Means** | **Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)** | |
| **1** | **2** | 299.232818 | 147.917950 | 450.547685 |



MomSmoke predictor variable has a significant effect on infant birth weight. Infants from non-smoking mom have more birth weight than infants from smoking moms. **Non-Smoking moms have larger mean value of weight than smoking mom**. The mean weight difference is significant as 95% confidence interval doesn’t include 0.

| **MomSmoke** | **Weight LSMEAN** | **95% Confidence Limits** | |
| --- | --- | --- | --- |
| **0** | 3323.765525 | 3245.115978 | 3402.415072 |
| **1** | 3056.881077 | 2896.656510 | 3217.105644 |

| **Least Squares Means for Effect MomSmoke** | | | | |
| --- | --- | --- | --- | --- |
| **i** | **j** | **Difference Between Means** | **Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)** | |
| **1** | **2** | 266.884448 | 104.090443 | 429.678454 |



## Exercise 5:

1. **What kind of ANOVA should we perform? And what additional assumption is required to perform this?**

**This dataset satisfy this assumption?**

We should perform **ANCOVA** to study the effect of categorical predictor variable (Trt) on dependent variable(final) by controlling/adjusting for the effect of covariate(initial).

We additionally need to check the **assumption of homogeneity of regression slope**.

For this we check the interaction of categorical predictor variable (Trt) and the continuous predictor variable (initial).

From Type III SS, we can conclude that there is **no significant interaction between Trt and Initial** variables (p-value 0.3602 is > 0.05 hence not significant at p < 0.05).

Since there is no interaction we can conclude that their slopes are equal and hence **assumption of homogeneity of slopes is met.**

| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Initial** | 1 | 68.52892317 | 68.52892317 | 241.81 | <.0001 |
| **Trt** | 4 | 1.69618688 | 0.42404672 | 1.50 | 0.2752 |
| **Initial\*Trt** | 4 | 1.38831410 | 0.34707852 | 1.22 | 0.3602 |

1. **Make a conclusion about location effect in final weight adjusting initial weight effect. How much of variation is explained by the model ?**

After performing ANCOVA, we find the location effect to be highly significant in final weight (p-value 0.0005 is < 0.05 hence significant at p < 0.05). Final weight of the oyster changes significantly with change in location (Trt).

| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Initial** | 1 | 156.0401767 | 156.0401767 | 517.38 | <.0001 |
| **Trt** | 4 | 12.0893593 | 3.0223398 | 10.02 | 0.0005 |

The model is also highly significant at p < 0.05 with a p-value < 0.0001. By R2 , we can conclude that the model explains 98.8% of variation in the final weight of the oysters. Hence the model is very useful and fits the data well.

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Model** | 5 | 354.4471767 | 70.8894353 | 235.05 | <.0001 |
| **Error** | 14 | 4.2223233 | 0.3015945 |  |  |
| **Corrected Total** | 19 | 358.6695000 |  |  |  |

| **R-Square** | **Coeff Var** | **Root MSE** | **Final Mean** |
| --- | --- | --- | --- |
| 0.988228 | 1.780438 | 0.549176 | 30.84500 |

1. **Comment on any significant group differences through post-hoc test. Comment on significant differences among 5 locations and specify which treatment is the most efficient in terms of oyster growth.**

From post-hoc test, the group differences between pairs 1-3, 1-4, 2-3, 2-4, 3-5 are significant as their Confidence intervals doesn’t include 0 which means their difference is not likely to b 0. Also the least square means table shows their respective p-values to be significant at p < 0.05.

Among all the treatments, the hot-bottom(Trt3) and hot-surface(Trt4) have significant effect on the final weight of the oysters. **Treatment 3 (hot-bottom) is most efficient** in terms of oyster growth then comes treatment 4 (hot-surface).

| **Least Squares Means for Effect Trt** | | | | |
| --- | --- | --- | --- | --- |
| **i** | **j** | **Difference Between Means** | **Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)** | |
| **1** | **2** | 0.035812 | -1.233043 | 1.304667 |
| **1** | **3** | -1.899217 | -3.326330 | -0.472104 |
| **1** | **4** | -1.351573 | -2.658254 | -0.044892 |
| **1** | **5** | -0.244459 | -2.040999 | 1.552080 |
| **2** | **3** | -1.935029 | -3.201670 | -0.668388 |
| **2** | **4** | -1.387385 | -2.602456 | -0.172314 |
| **2** | **5** | -0.280271 | -1.816097 | 1.255554 |
| **3** | **4** | 0.547644 | -0.690652 | 1.785940 |
| **3** | **5** | 1.654758 | 0.316719 | 2.992796 |
| **4** | **5** | 1.107114 | -0.362790 | 2.577017 |

| **Trt** | **Final LSMEAN** | **LSMEAN Number** |
| --- | --- | --- |
| **1** | 30.1531125 | 1 |
| **2** | 30.1173006 | 2 |
| **3** | 32.0523296 | 3 |
| **4** | 31.5046854 | 4 |
| **5** | 30.3975719 | 5 |

| **Least Squares Means for effect Trt Pr > |t| for H0: LSMean(i)=LSMean(j)  Dependent Variable: Final** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **i/j** | **1** | **2** | **3** | **4** | **5** |
| **1** |  | 1.0000 | 0.0074 | 0.0411 | 0.9925 |
| **2** | 1.0000 |  | 0.0024 | 0.0222 | 0.9776 |
| **3** | 0.0074 | 0.0024 |  | 0.6504 | 0.0128 |
| **4** | 0.0411 | 0.0222 | 0.6504 |  | 0.1875 |
| **5** | 0.9925 | 0.9776 | 0.0128 | 0.1875 |  |

| **Trt** | **Final LSMEAN** | **95% Confidence Limits** | |
| --- | --- | --- | --- |
| **1** | 30.153113 | 29.436931 | 30.869294 |
| **2** | 30.117301 | 29.510894 | 30.723707 |
| **3** | 32.052330 | 31.452584 | 32.652075 |
| **4** | 31.504685 | 30.911849 | 32.097522 |
| **5** | 30.397572 | 29.620733 | 31.174411 |

