**Define :-**

A magazine does a survey and wants to publish results. The following data represent business startup costs (thousands of dollars) for shops, gathered thru a survey. One of the Questions to be answered is that is it equally expensive to start any of the businesses explored below or are some businesses lighter on the pocket?

Description of variables:-

* X1 = startup costs for pizza
* X2 = startup costs for baker/donuts
* X3 = startup costs for shoe stores
* X4 = startup costs for gift shops
* X5 = startup costs for pet stores

What is the conclusion that the article should publish?

/\*D – Define the business Y – check if the distribution of startup costs are similar across lines of business\*/

/\* IMPORT DATA\*/

**DATA** WORK.STARTUP;

LENGTH

X1 **8**

X2 **8**

X3 **8**

X4 **8**

X5 **8** ;

FORMAT

X1 BEST3.

X2 BEST3.

X3 BEST3.

X4 BEST3.

X5 BEST3. ;

INFORMAT

X1 BEST3.

X2 BEST3.

X3 BEST3.

X4 BEST3.

X5 BEST3. ;

INFILE '/home/subhashini1/my\_content/startupcost.csv'

DLM=','

MISSOVER

DSD ;

INPUT

X1 : ?? BEST3.

X2 : ?? BEST3.

X3 : ?? BEST3.

X4 : ?? BEST3.

X5 : ?? BEST3. ;

**RUN**;

/\*UNDERSTAND THE DATA \*/

**PROC** **CONTENTS** DATA = WORK.STARTUP; **RUN** ;

**PROC** **PRINT** DATA=WORK.STARTUP (OBS= **10**) ; **RUN** ;

**PROC** **UNIVARIATE** DATA= WORK.STARTUP NORMAL PLOT;

VAR X1 X2 X3 X4;

HISTOGRAM X2 X3 X4 / NORMAL ; **RUN** ;

/\* STRUCTURE THE DATA FOR T TEST X1 VS X2 \*/

**DATA** WORK.STARTUP\_X1;

SET WORK.STARTUP;

KEEP X1; **RUN** ;

**DATA** WORK.STARTUP\_X1;

SET WORK.STARTUP\_X1;

SEGMENT = **1**;

RENAME X1 = VALUE;**RUN** ;

**PROC** **PRINT** DATA=WORK.STARTUP\_X1 (OBS=**10**) ; **RUN** ;

**DATA** WORK.STARTUP\_X2;

SET WORK.STARTUP;

KEEP X2;

**RUN** ;

**DATA** WORK.STARTUP\_X2;

SET WORK.STARTUP\_X2;

SEGMENT = **2**;

RENAME X2 = VALUE;

**RUN** ;

**PROC** **PRINT** DATA=WORK.STARTUP\_X2 (OBS=**10**) ; **RUN** ;

**DATA** TOTAL\_X1\_X2;

SET WORK.STARTUP\_X1 WORK.STARTUP\_X2;**RUN** ;

**PROC** **SORT** DATA= TOTAL\_X1\_X2;

BY SEGMENT; **RUN** ;

/\* UNDERSTAND MISSING VALUES \*/

**PROC** **MEANS** DATA=TOTAL\_X1\_X2 N NMISS MEAN STDDEV MEDIAN MIN MAX;

CLASS SEGMENT ; **RUN** ;

/\* REMOVE MISSING VALUES\*/

**DATA** FINAL\_X1\_X2;

SET TOTAL\_X1\_X2;

WHERE VALUE NE **.**; **RUN** ;

**PROC** **MEANS** DATA=FINAL\_X1\_X2 N NMISS MEAN STDDEV MEDIAN MIN MAX;

CLASS SEGMENT ; **RUN** ;

/\* CHECK NORMALITY \*/

**PROC** **UNIVARIATE** DATA=FINAL\_X1\_X2 NORMAL;

QQPLOT VALUE / NORMAL (mu=est sigma=est color=red l=**1**);

BY SEGMENT; **RUN** ;

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | |  | | --- | | **The UNIVARIATE Procedure** | | **SEGMENT=1** | | | |  | | --- | | img3.png | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | |  | | --- | | **The UNIVARIATE Procedure Variable:  VALUE** | | **SEGMENT=2** | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | | **Moments** | | | | | --- | --- | --- | --- | | N | 22 | Sum Weights | 22 | | Mean | 92.0909091 | Sum Observations | 2026 | | Std Deviation | 37.956001 | Variance | 1440.65801 | | Skewness | 0.47019131 | Kurtosis | -0.6371584 | | Uncorrected SS | 216830 | Corrected SS | 30253.8182 | | Coeff Variation | 41.2157957 | Std Error Mean | 8.09224659 | | | | **Basic Statistical Measures** | | | | | --- | --- | --- | --- | | **Location** | | **Variability** | | | Mean | 92.09091 | Std Deviation | 37.95600 | | Median | 87.00000 | Variance | 1441 | | Mode | 40.00000 | Range | 120.00000 | |  |  | Interquartile Range | 60.00000 | | | | |  |  | | --- | --- | | |  | | --- | | **Note: The mode displayed is the smallest of 11 modes with a count of 2.** | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | | **Tests for Location: Mu0=0** | | | | | | --- | --- | --- | --- | --- | | **Test** | **Statistic** | | **p Value** | | | Student's t | t | 11.38014 | Pr > |t| | <.0001 | | Sign | M | 11 | Pr >= |M| | <.0001 | | Signed Rank | S | 126.5 | Pr >= |S| | <.0001 | | | | **Tests for Normality** | | | | | | --- | --- | --- | --- | --- | | **Test** | **Statistic** | | **p Value** | | | Shapiro-Wilk | W | 0.921498 | Pr < W | 0.0817 | | Kolmogorov-Smirnov | D | 0.158329 | Pr > D | >0.1500 | | Cramer-von Mises | W-Sq | 0.084843 | Pr > W-Sq | 0.1749 | | Anderson-Darling | A-Sq | 0.572258 | Pr > A-Sq | 0.1261 | | | | **Quantiles (Definition 5)** | | | --- | --- | | **Level** | **Quantile** | | 100% Max | 160 | | 99% | 160 | | 95% | 160 | | 90% | 150 | | 75% Q3 | 120 | | 50% Median | 87 | | 25% Q1 | 60 | | 10% | 45 | | 5% | 40 | | 1% | 40 | | 0% Min | 40 | | | | **Extreme Observations** | | | | | --- | --- | --- | --- | | **Lowest** | | **Highest** | | | **Value** | **Obs** | **Value** | **Obs** | | 40 | 26 | 120 | 27 | | 40 | 15 | 150 | 14 | | 45 | 31 | 150 | 25 | | 45 | 20 | 160 | 18 | | 60 | 30 | 160 | 29 | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | |  | | --- | | **The UNIVARIATE Procedure** | | **SEGMENT=2** | | | |  | | --- | | img4.png | | |

Conclusion :- The QQplots appear linear. The NULL HYPOTHESIS for Normality test is that there is there is no significant departure frim normality . Since the p value is greater than .05 , we cannot reject the null.

Notes :- If the sample size is over 2000, the Kolmgorov test should be used. If the sample size is less than 2000, the Shapiro test is better.

**Problem for you :- Please do this for the other variables in the dataset and conclude on the distribution**

/\* RUN T TEST FOR 2 SAMPLES, INDEPENDENT GROUP \*/

**PROC** **TTEST** DATA=FINAL\_X1\_X2;

CLASS SEGMENT;

VAR VALUE; **RUN** ;

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  | | --- | --- | | |  | | --- | | **The TTEST Procedure  Variable:  VALUE** | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | | **SEGMENT** | **N** | **Mean** | **Std Dev** | **Std Err** | **Minimum** | **Maximum** | | --- | --- | --- | --- | --- | --- | --- | | 1 | 13 | 83.0000 | 34.1345 | 9.4672 | 35.0000 | 140.0 | | 2 | 22 | 92.0909 | 37.9560 | 8.0922 | 40.0000 | 160.0 | | Diff (1-2) |  | -9.0909 | 36.6126 | 12.8080 |  |  | | | | **SEGMENT** | **Method** | **Mean** | **95% CL Mean** | | **Std Dev** | **95% CL Std Dev** | | | --- | --- | --- | --- | --- | --- | --- | --- | | 1 |  | 83.0000 | 62.3727 | 103.6 | 34.1345 | 24.4774 | 56.3471 | | 2 |  | 92.0909 | 75.2622 | 108.9 | 37.9560 | 29.2015 | 54.2416 | | Diff (1-2) | Pooled | -9.0909 | -35.1490 | 16.9671 | 36.6126 | 29.5308 | 48.1923 | | Diff (1-2) | Satterthwaite | -9.0909 | -34.6218 | 16.4400 |  |  |  | | | | **Method** | **Variances** | **DF** | **t Value** | **Pr > |t|** | | --- | --- | --- | --- | --- | | Pooled | Equal | 33 | -0.71 | 0.4828 | | Satterthwaite | Unequal | 27.54 | -0.73 | 0.4716 | | | | **Equality of Variances** | | | | | | --- | --- | --- | --- | --- | | **Method** | **Num DF** | **Den DF** | **F Value** | **Pr > F** | | Folded F | 21 | 12 | 1.24 | 0.7210 | | | | |

Conclusion :- Since p value is greater than .05, you CANNOT reject the null that mean1 = mean 2 .

**Problem for you :- Please do this for the other variables in the dataset and conclude on the hypothesis test**

/\* RUN ANOVA (IN ANOVA THE ASSUMPTIONS ARE TESTED AFTER RUNNING THE MODEL)\*/

**PROC** **ANOVA** DATA=WORK.FINAL\_X1\_X2;

CLASS SEGMENT;

MODEL VALUE = SEGMENT ;

MEANS SEGMENT /HOVTEST=LEVENE;**RUN**;

|  |
| --- |
|  |
|  |
|  |
| |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  | | --- | --- | | |  | | --- | | **The ANOVA Procedure** | | | |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | | **Class Level Information** | | | | --- | --- | --- | | **Class** | **Levels** | **Values** | | SEGMENT | 2 | 1 2 | | | | Number of Observations Read | 35 | | --- | --- | | Number of Observations Used | 35 | | | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  | | --- | --- | | |  | | --- | | **The ANOVA Procedure  Dependent Variable: VALUE** | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | | **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** | | --- | --- | --- | --- | --- | --- | | Model | 1 | 675.32468 | 675.32468 | 0.50 | 0.4828 | | Error | 33 | 44235.81818 | 1340.47934 |  |  | | Corrected Total | 34 | 44911.14286 |  |  |  | | | | **R-Square** | **Coeff Var** | **Root MSE** | **VALUE Mean** | | --- | --- | --- | --- | | 0.015037 | 41.27019 | 36.61256 | 88.71429 | | | | **Source** | **DF** | **Anova SS** | **Mean Square** | **F Value** | **Pr > F** | | --- | --- | --- | --- | --- | --- | | SEGMENT | 1 | 675.3246753 | 675.3246753 | 0.50 | 0.4828 | | | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  | | --- | --- | | |  | | --- | | **The ANOVA Procedure** | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | | **Levene's Test for Homogeneity of VALUE Variance ANOVA of Squared Deviations from Group Means** | | | | | | | --- | --- | --- | --- | --- | --- | | **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** | | SEGMENT | 1 | 733641 | 733641 | 0.37 | 0.5455 | | Error | 33 | 64871914 | 1965816 |  |  | | | | **Welch's ANOVA for VALUE** | | | | | --- | --- | --- | --- | | **Source** | **DF** | **F Value** | **Pr > F** | | SEGMENT | 1.0000 | 0.53 | 0.4716 | | Error | 27.5400 |  |  | | | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  | | --- | --- | | |  | | --- | | **The ANOVA Procedure** | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | | **Level of SEGMENT** | **N** | **VALUE** | | | --- | --- | --- | --- | | **Mean** | **Std Dev** | | 1 | 13 | 83.0000000 | 34.1345377 | | 2 | 22 | 92.0909091 | 37.9560010 | | | | | |

Conclusion :- We CANNOT reject the null hypothesis that Mean 1 = Mean 2 at 95% confidence (this is the default values)

Notes :-  The assumption of equal variances (homogeneity of variances) can be checked with "hovtest" option where we have used Levene’s test . Here the p value is .5455 and therefore we cannot reject the null

The Welch test is for testing the hypothesis assuming No Homogeneity and the p value for this is .4716. Therefore we cannot accept the null .

**Problem for you :- Please do this for the other variables in the dataset and conclude on all the 5 variables together. For doing this you will need to create a dataset with all 5 variables (X1- X5)**