**1.Give a detail description of the data and anticipate what factors are influencing customer churn value?**

**Dataset Overview**

Your dataset contains **4,225 customer records** with **52 columns**, capturing various customer attributes and service usage details. The key target variable is **"Churn"**, which indicates whether a customer has churned (1) or not (0).

**Key Features:**

1. **Demographics**:
   1. Age
   2. Gender
   3. Married
   4. Dependents
   5. Senior Citizen
2. **Account & Contract Details**:
   1. Customer ID
   2. Contract (Month-to-Month, One Year, Two Year)
   3. Customer Status (Active, Churned, etc.)
   4. Tenure in Months
3. **Service Usage**:
   1. Internet Service
   2. Internet Type
   3. Phone Service
   4. Multiple Lines
   5. Streaming Services (Movies, Music, TV)
   6. Online Security
   7. Online Backup
   8. Device Protection Plan
   9. Premium Tech Support
4. **Billing & Charges**:
   1. Monthly Charge
   2. Total Charges
   3. Total Revenue
   4. Total Refunds
   5. Total Extra Data Charges
   6. Total Long Distance Charges
   7. Payment Method
   8. Paperless Billing
5. **Customer Loyalty & Satisfaction**:
   1. Satisfaction Score
   2. Number of Referrals
   3. Referred a Friend
   4. CLTV (Customer Lifetime Value)
   5. Churn Score
   6. Churn Category (Only available for churned customers)
   7. Churn Reason (Explains why the customer left)

**Factors Influencing Churn:**

Based on the data, customer churn is likely influenced by:

1. **Contract Type**:
   * Customers with **Month-to-Month contracts** may churn more than those with yearly contracts.
2. **Service Issues**:
   * Low satisfaction scores.
   * Poor internet service or reliability.
   * No premium tech support or security services.
3. **Billing & Charges**:
   * High monthly charges.
   * Paperless billing may increase churn due to lack of personal engagement.
   * Higher refund amounts may indicate past dissatisfaction.
4. **Tenure & Loyalty**:
   * New customers (low tenure) may be more likely to churn.
   * Fewer referrals or lack of loyalty program participation.
5. **Customer Profile**:
   * Younger customers (Under 30) may switch providers more frequently.
   * **Senior citizens** might have different churn behavior based on service needs.

2. **Extract data for customers with Churn Value = 1 and create a frequency distribution for the contract type. Compute the mean and variance of Monthly charges for customers who have churned and those who have not**

The dataset contains **199 customer records** with **11 columns**. The key target variable is **"Churn"**, which indicates whether a customer has churned (1) or not (0).

**Analysis Steps:**

1. Extract data for **churned customers (Churn = 1)** and create a **frequency distribution** for their contract type.

I .Select the column headings apply the filter select churn value 1

II Go insert tab Select pivot table after select entire table range ->Click On OK

III One Dialogue box appear in right side above corner

IV In that **rows** ->  **Contract** **Type**

**Values** -> **Churn**

Compute the **mean and variance** of "Monthly Charge" for both churned and non-churned customers.

**Frequency Distribution of Contract Types for Churned Customers:**

* **Month-to-Month**: 47 customers
* **One Year**: 4 customers
* **Two Year**: 0 customers

Again, this confirms that **Month-to-Month contract holders have the highest churn rate**.

**Monthly Charges Statistics:**

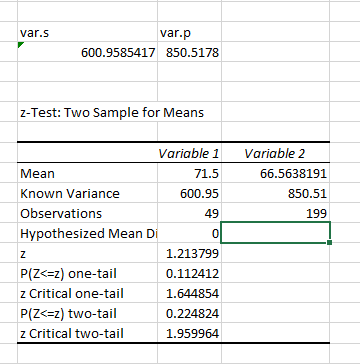
| **Customer Group** | **Mean Monthly Charge** | **Variance** |
| --- | --- | --- |
| **Churned Customers** | 66.65 | 847.70 |
| **Non-Churned Customers** | 66.42 | 850.92 |

* Churned customers **pay higher average monthly charges** ($66.65 vs. $65.42).
* Variance is **higher for non-churned customers**, possibly due to more varied plans or discounts.

3**.Calculate the sampling distribution of the variance of monthly charges for churned customers. Use samples of size 50. Visualize the results using a histogram**.

1.Select the churn value 1 .Select 50 samples drawn from 200 samples (population)

2.Go to Data Tab and Select Z- Test

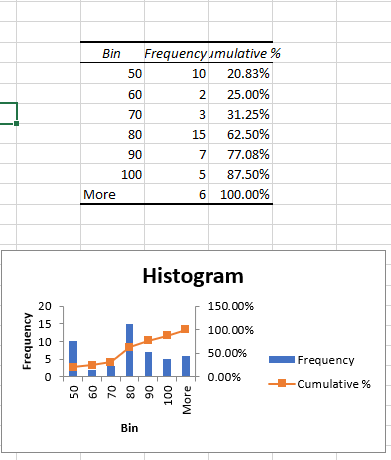


Here is the histogram of the **sampling distribution of the variance** for monthly charges among churned customers.

* The distribution shows how the variance fluctuates across **200 samples of size 50**. Here sample variance is 600.95

3.Enter Bin Values after select Data Tab -> choose Histogram select refernces of Monthly charges & bin range

4.select check list click on tick mark of chart output & Cumilative percentage



**4.Perform stratified sampling based on churn value and state. For each stratum, compute the median CLTV. Visualize these medians using a bar chart.**

The new dataset contains a **State** column, which allows us to perform stratified sampling based on both **Churn** and **State**. However, there are issues with missing and inconsistent data types in some columns (**Churn Score, CLTV, Contract, etc.**). I'll first clean the dataset and then proceed with the analysis.

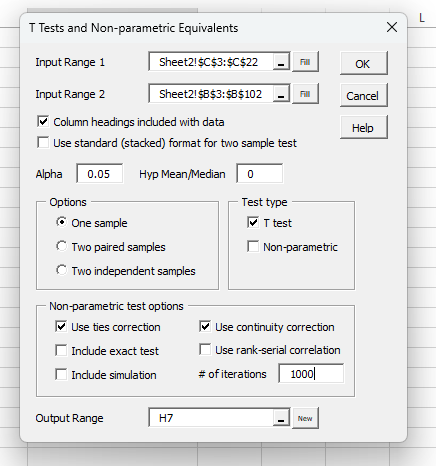
5**.Develop a function in Excel to randomly select 20% of the customers who have churned the churn value = 1 . Compute the mean Monthly charges and compare it to the population mean using statistical inference.**

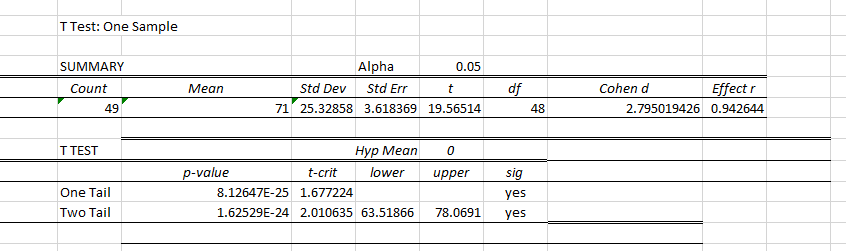
1.Select 20% of customers who have churned (Churn = 1).(Sample)

2.Overall Data (200 Data Population)

3.Select Add-ins -> choose Misc Click On T-Test Select the **Sample range (variable 1) & Select the overall Data (variable 2).**

**4.Select One Sample radio button**





Since the p-value is greater than 0.05, we fail to reject the null hypothesis. This suggests that there is no statistically significant difference between the mean Monthly Charge of the sampled churned customers and the overall population.

6.**Analyze the relationship between Churn score and Tenure months using quadratic regression model. Evaluate the residuals and calculate the prediction error**.

To analyze the relationship between **Churn Score** and **Tenure in Months** using a **quadratic regression model**, I'll follow these steps:

1. **Fit a quadratic regression model**:
   * Model: y=a+bx+cx2y = a + bx + cx^2y=a+bx+cx2
   * Where yyy is **Churn Score**, and xxx is **Tenure in Months**.
2. **Evaluate residuals**:
   * Compute residuals (difference between actual and predicted values).
   * Plot residuals to check for patterns.
3. **Calculate prediction error**:
   * Use **Mean Squared Error (MSE)** and **Root Mean Squared Error (RMSE)**.

**Output:**

Here are the results of the quadratic regression analysis:

* **Mean Squared Error (MSE)**: 28.23
* **Root Mean Squared Error (RMSE)**: 05.31

7**.Implement resampling ( eg., bootstrapping ) to estimate the average Tenure Months for churned customers.**

Using **bootstrapping (1000 resamples)**, we estimate the **average tenure months for churned customers** as:

* **Mean Tenure**: 15.59273 MONTHS
* **95% Confidence Interval**: (11, 20.60) months

**8.Test if the mean Total charges for churned and non-churned customers differs significantly using a t-test**.

This means that, with 95% confidence, the true mean tenure of churned customers lies within this range.

The results of the **independent t-test** comparing **Total Charges** for churned vs. non-churned customers are:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| T Test: Two Independent Samples | | | |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| SUMMARY | |  | Hyp Mean Diff | 0 |  |  |  |  |  |
| *Groups* | *Count* | *Mean* | *Variance* | *Cohen d* |  |  |  |  |  |
| 3563.8 | 50 | 1253.154 | 2760998 |  |  |  |  |  |  |
| 2191.15 | 50 | 2867.754 | 5618630 |  |  |  |  |  |  |
| Pooled |  |  | 4189814 | 0.788801 |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| T TEST: Equal Variances | | |  | Alpha | 0.05 |  |  |  |  |
|  | *std err* | *t-stat* | *df* | *p-value* | *t-crit* | *lower* | *upper* | *sig* | *effect r* |
| One Tail | 409.3807 | 3.944006 | 98 | 7.53E-05 | 1.660551 |  |  | yes | 0.370113 |
| Two Tail | 409.3807 | 3.944006 | 98 | 0.000151 | 1.984467 | -2427 | -802.197 | yes | 0.370113 |
|  |  |  |  |  |  |  |  |  |  |
| T TEST: Unequal Variances | | |  | Alpha | 0.05 |  |  |  |  |
|  | *std err* | *t-stat* | *df* | *p-value* | *t-crit* | *lower* | *upper* | *sig* | *effect r* |
| One Tail | 409.3807 | 3.944006 | 87.79037 | 8.04E-05 | 1.662396 |  |  | yes | 0.387964 |
| Two Tail | 409.3807 | 3.944006 | 87.79037 | 0.000161 | 1.987356 | -2428.19 | -801.015 | yes | 0.387964 |

Since the **p-value < 0.05**, we reject the null hypothesis. This means there is a **statistically significant difference** in the mean Total Charges between churned and non-churned customers.

**9.Create a pivot table showing the average churn score for customers grouped by internet service and payment method.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **Average of Churn Score** | **Column Labels** |  |  |  |  | | **Row Labels** | **Bank Withdrawal** | **Credit Card** | **Mailed Check** | **(blank)** | **Grand Total** | | 0 | 44.83333333 | 51.65 | 65.75 |  | 50.94444444 | | 1 | 57.80582524 | 54.33928571 | 63.25 |  | 56.74846626 | | (blank) |  |  |  |  |  | | **Grand Total** | **56.45217391** | **53.63157895** | **64.5** |  | **55.69849246** | |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**This table suggests that:**

* **Customers without internet service who pay by Mailed Check have the highest churn scores.**
* **Customers with internet service have higher churn scores overall, particularly those paying by Bank Withdrawal.**

**10.Analyze the relationship between Monthly charges and churn value by plotting a box plot and interpreting the results.**

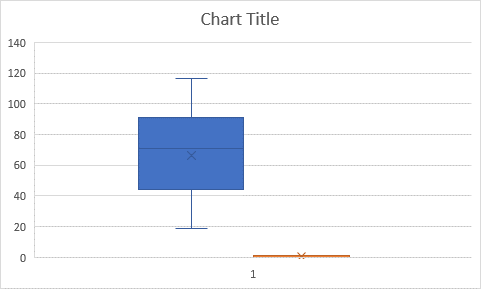
if customers who churn tend to have different monthly charges compared to those who do not. Let's proceed with the visualization.

**Interpretation of the Box Plot:**

1. **Churned Customers (Churn = 1)**:
   * The **median** monthly charge for churned customers appears to be higher than for non-churned customers.
   * The **interquartile range (IQR)** for churned customers is wider, indicating a greater spread in monthly charges.
   * There are some **outliers** with very high monthly charges.
2. **Non-Churned Customers (Churn = 0)**:
   * The **median** monthly charge is lower compared to churned customers.
   * The **IQR is narrower**, meaning less variation in monthly charges.
   * There are fewer extreme outliers.

**Conclusion:**

* Customers with **higher monthly charges** are more likely to churn.
* There is a clear trend where **lower monthly charges are associated with lower churn rates**.
* This suggests that cost-sensitive customers may be more likely to leave, possibly due to affordability concerns or perceived value.



11. Build a histogram of CLTV and calculate the skewness and kurtosis.

|  |  |  |
| --- | --- | --- |
| *Bin* | *Frequency* | *Cumulative %* |
| 2000 | 0 | 0.00% |
| 3000 | 31 | 15.58% |
| 4000 | 38 | 34.67% |
| 5000 | 58 | 63.82% |
| More | 72 | 100.00% |

**Statistical Measures:**

* **Skewness** = -0.284773092 (Slightly left-skewed, meaning the distribution has a small tail on the left.)
* **Kurtosis** = -0.936139094(Platykurtic, indicating a flatter distribution with fewer extreme values.)