MA5701: Statistical Methods
Project Report

Statistical Analysis of Telco Churn Data

By

Group 8

Dhanush Biligiri N H

Inderpreet Singh Juneja

Rachana Tanneeru

1. Introduction:

The Telco industry is highly competitive, with numerous players vying for customers. In this context, customer retention is critical for the success of any telecom company. Retaining existing customers is not only more cost-effective than acquiring new ones but also leads to increased customer loyalty and positive word-of-mouth.

The Telco Customer Churn dataset is one such dataset that contains information about customers of a telecom company, including their demographic information, account information, and usage information. The dataset is used for analyzing monthly charges variation and understanding the factors that contribute to it. By analyzing this dataset, telecom companies can gain valuable insights into the factors that drive customer churn and take measures to retain their customers.

Few of the attributes comprising this dataset are:

CustomerID: a unique identifier for each customer

Gender: the gender of the customer (Male/Female)

InternetService: the type of internet service the customer has (DSL/Fiber optic/No)

Contract: the type of contract the customer has (Month-to-month/One year/Two year)

PaymentMethod: the customer's payment method (Electronic check/Mailed check/Bank transfer (automatic)/Credit card (automatic))

MonthlyCharges: the amount charged to the customer monthly.

TotalCharges: the total amount charged to the customer over their tenure with the company.

Among these diverse factors, the following variables have been used for analysis:

Categorical variable: Internet Service, Contract, Payment Method

Numerical variable: Monthly Charges, Total Charges

The selection of these items was made based on exploratory data analyses that identified highly influential factors.

Initial Hypotheses:

1. It is expected that changes in the Internet Service type might have an impact on the monthly charges.
2. Similarly, a change in contract type might also affect the monthly charges.

2. Method:

Sampling unit: In the Telco Customer Churn dataset, the sampling units consist of individual customers of the telecommunications company. The dataset could include data on either a representative subset of customers or the entire customer base of the company.

The analysis carried out is in the form of an observational study. Observational studies are research designs where the researcher records and observes data without interfering or manipulating the subjects under study. The variables being measured are not controlled by the researcher, and the researcher simply records the natural characteristics or behavior of the participants. Although these studies do not establish direct causation, they can still offer valuable information about the association between variables and provide a basis for further research. It is important to be cautious of confounding variables that could introduce bias since observational studies lack randomization and may not control all possible sources of bias.

Website referred for dataset: Telco customer churn

2.1 Exploratory Data Analyses:

Our data contained some null values; hence those null values were omitted due to which total dimension of the data has become 7032 x 21 from 7043 x 23. The data was checked for duplicates rows as well, but none were found.

Apart from the actual variables used for data analyses, few extra relationship analyses have been performed for better understanding of the churn pattern. Based on these, the categorical and numerical variables mentioned earlier were chosen.

2.2 Plots used for EDA:

Figure 1: Histogram for Churn vs Contract Type, Payment method

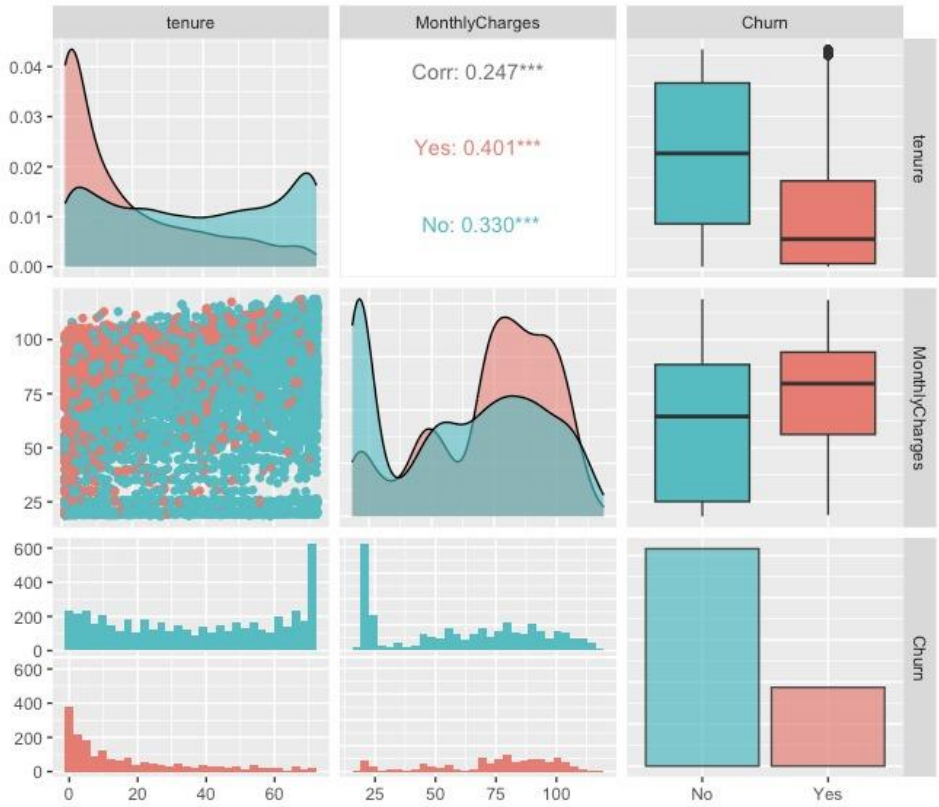

Figure 2: Correlation plots for Tenure, Monthly Charges, Churn

There is a moderate negative correlation between tenure and churn which suggests that customers who have been with the company for a longer period are less likely to churn.

There is a positive correlation between monthly charges and churn (0.19), which suggests that customers who pay higher monthly charges are more likely to churn.

There is a weak positive correlation between tenure and monthly charges (0.25), which suggests that customers who have been with the company for a longer period tend to have higher monthly charges.

3. Results:

The concern of these analyses is to make inferences from the statistical analysis of likely influential factors and establish with some confidence, whether these factors are responsible for change in the churn counts of the company.

3.1 One-way ANOVA:

Dependent variable: Monthly Charges (in $)

Independent variable: Internet Service (DSL/Fiber optic/No)

No. of observations: 7032

For initial analyses, a set of parallel boxplots have been created as shown:

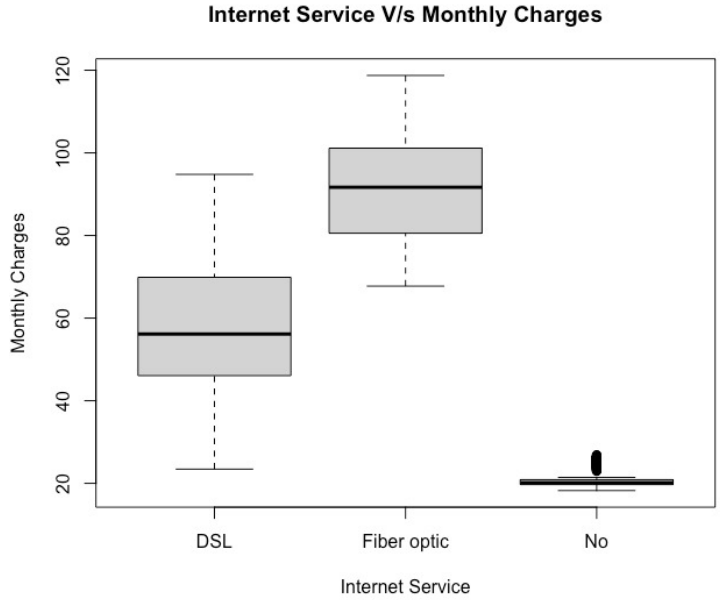

Figure 3: Parallel Boxplots of Independent vs Dependent variable

The nature of figure 3 suggests that there is a significant increase in the monthly charges paid by the customers when the internet service switches from DSL to Fiber optic. However, third level in this category is “No”, representing the number of customers who did not opt for an internet service.

Attributes	DSL	Fiber Optic	No
Sample Size	2416	3091	1525
Mean	58.08802	91.50013	21.07628
Median	56.150	91.675	20.150
Standard Deviation	16.266167	12.663039	2.161599

Table 1: Sample Statistics

Table 1 displays the sample size, sample mean, sample median, and sample standard deviation of monthly charges according to each level of internet service level.

ANOVA is conducted with the following hypothesis:

Null Hypothesis (H0): All population means are equal.

Alternative Hypothesis (Ha): Atleast one population mean is different from others.

Confidence level assumed = 95%

```
Analysis of Variance Table

Response: MonthlyCharges
                Df  Sum Sq Mean Sq F value    Pr(>F)
InternetService  2 5221852 2610926   16065 < 2.2e-16 ***
Residuals       7029 1142369     163
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4: One-way ANOVA table

From the ANOVA table. P-value is very small as compared to the significance level, $\alpha = 0.05$. Hence, we reject the null hypothesis and state that there is strong enough evidence to say that the atleast one population means is different than others.

Checking the validity of Assumptions of one-way ANOVA:

1. Normality of residuals:
   This can be checked using a Normal Q-Q plot and a Shapiro-Wilk test applied to the residuals.

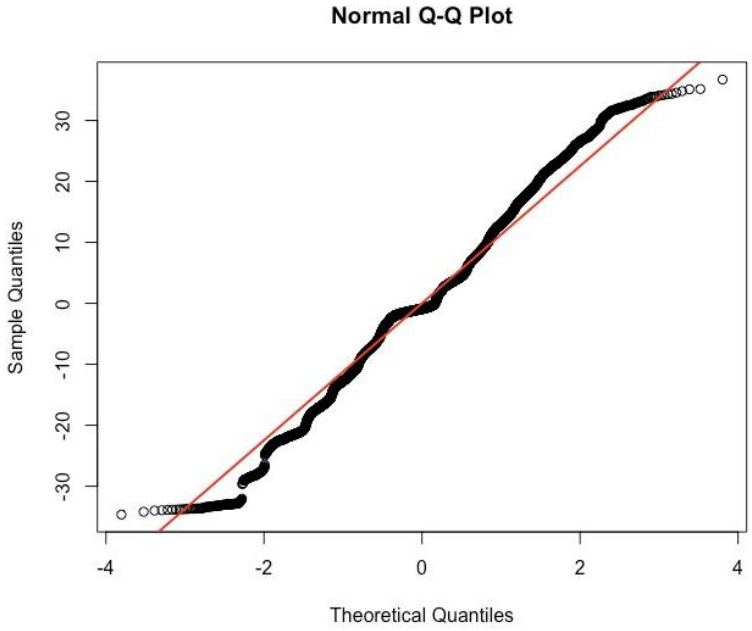

Figure 4: Normal Q-Q plot of residuals

The residuals of 5000+ samples appear to be following a normal nature, but it is based on this graph. Furthermore, the p-value of Shapiro-wilk test is calculated and is found to be smaller than $\alpha = 0.05$ which means that the residuals are not normally distributed. This test, in case of this data set is not completely reliable to its limitations of computing

only 5000 residuals at a time, hence we go ahead with Anderson-Darling test and found similar results from that as well. The results from these tests are as follows:

```
        Shapiro-Wilk normality test

data:  data$resid1[1:5000]
W = 0.98932, p-value < 2.2e-16
```

```
        Anderson-Darling normality test

data:  data$resid1
A = 34.627, p-value < 2.2e-16
```

Figure 5: Normality tests

At last, we can state that the normality assumption has been violated according to these tests, however based on the nature of Normal Q-Q plot, the data still seems to be fairly normal.

2. Assumption of Homoscedasticity:
   Residuals Vs Fitted values plot has been plotted below which shows an obvious pattern indicating that this graph might not be a good fit for the data.

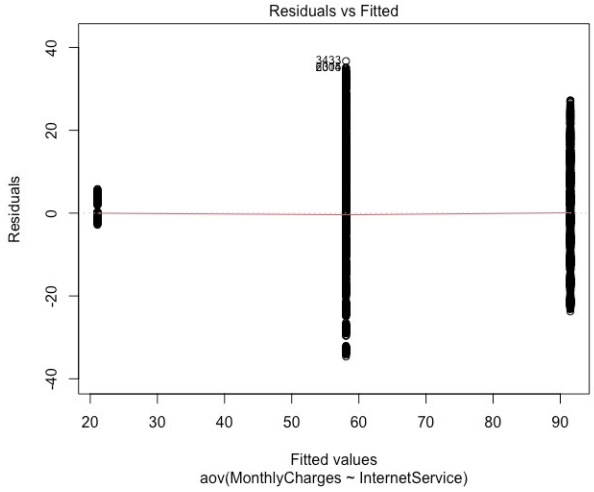

Figure 6: Residuals Vs Fitted values

```
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    2  1314.8 < 2.2e-16 ***
      7029
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 7: Levene’s Test for equal variances

From the output of Levene’s test (p-value < $\alpha = 0.05$) and nature of Residuals Vs Fitted values plot, it is concluded that the assumption of Homoscedasticity has been violated.

For us to perform One-way ANOVA, certain assumptions must be met -

- IID: we have k random samples collected from their respective populations.
- Independent Samples: the k samples are collected independently of one another.

- Normal Pop.: the k populations are all normally distributed.
- Homogeneity of Var.: the k populations possess the property of homoscedasticity (they all have the same population variance)

From the Q-Q plot we conclude that the datapoints are normal. However, from the Levene's test we conclude that the data violates the assumption of homoscedasticity. Since the sample size for each population group is very large, the data can be declared to be normal in terms of robustness to the violation of normality assumption. This is the reason we select Welch's ANOVA as the assumption of the test matches with the assumptions of our data.

Welch's ANOVA:

Null hypothesis (H0): The population means of the two groups are equal

Alternative hypothesis (Ha): The population means of the two groups are not equal

With all the assumptions of Welch's ANOVA test confirmed, test is conducted with following output:

```
        One-way analysis of means (not assuming equal variances)

data:  MonthlyCharges and InternetService
F = 50050, num df = 2.0, denom df = 3887.1, p-value < 2.2e-16
```

Figure 7: Welch's ANOVA test output

With a moderate significance level of 0.05, it can be concluded that the null hypothesis can be rejected confirming that there is atleast one population mean which is significantly different than others.

Now that we know atleast one mean is different, it is suitable to conduct a post hoc multiple comparison procedure i.e., pairwise.t.test to compare all pairs of means and report their p-values which will be compared with a significance level of $\alpha = 0.05$.

```
             DSL    Fiber optic
Fiber optic <2e-16 -
No          <2e-16 <2e-16

P value adjustment method: holm
```

Figure 8: Pairwise t-test output

From the table above, it is evident that all the p-values are smaller than significance level and conclude that all the means are significantly different from each other.

Based on this output and parallel boxplot from Figure 3, we conclude that 'Fiber optic' has significantly greater mean of Monthly Charges than both 'DSL' and 'No' internet services.

3.2 One-way ANOVA:

Dependent variable: Monthly Charges (in $)

Independent variable: Contract (Month-to-month/One-year/Two-year)

No. of observations: 7032

For initial analyses, a set of parallel boxplots have been created as shown:

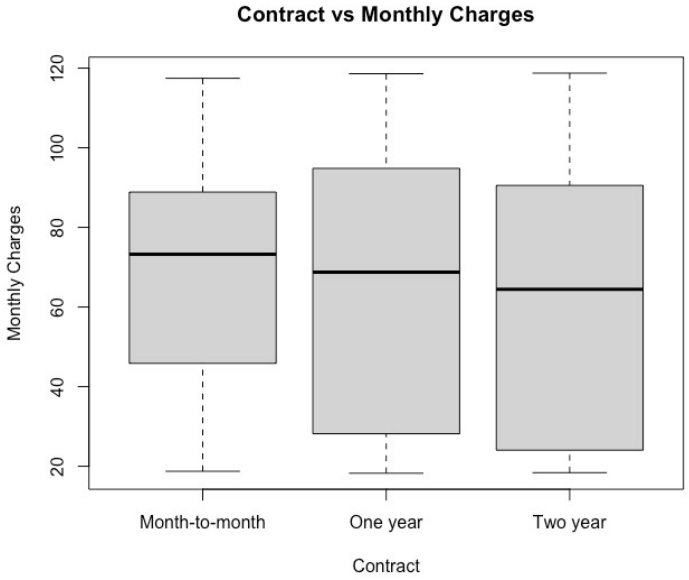

Figure 9: Parallel Boxplots of Independent vs Dependent variable

The nature of figure 9 suggests that monthly charges paid by the customers are somewhat similar for all three types of contracts.

Attributes	Month -to -Month	One Year	Two Years
Sample Size	3870	1468	1694
Mean	66.39849	65.04861	60.77041
Median	73.25	68.75	64.35
Standard Deviation	26.92660	31.84054	34.67887

Table 2: Sample Statistics

Table 2 displays the sample size, sample mean, sample median, and sample standard deviation of monthly charges according to each of contract.

ANOVA is conducted with the following hypothesis:

Null Hypothesis (H0): All population means are equal.

Alternative Hypothesis (Ha): Atleast one population mean is different from others.

Confidence level assumed = 95%

```
Analysis of Variance Table

Response: MonthlyCharges
            Df  Sum Sq Mean Sq F value    Pr(>F)
Contract     2   36009 18004.7  19.999 2.185e-09 ***
Residuals 7029 6328211   900.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 10: One-way ANOVA table

From the ANOVA table. P-value is very small as compared to the significance level, $\alpha = 0.05$. Hence, we reject the null hypothesis and state that there is strong enough evidence to say that the atleast one population mean is different than others.

Checking the validity of Assumptions of one-way ANOVA:

1. Normality of residuals:
   This can be checked using a Normal Q-Q plot and a Shapiro-Wilk test applied to the residuals.

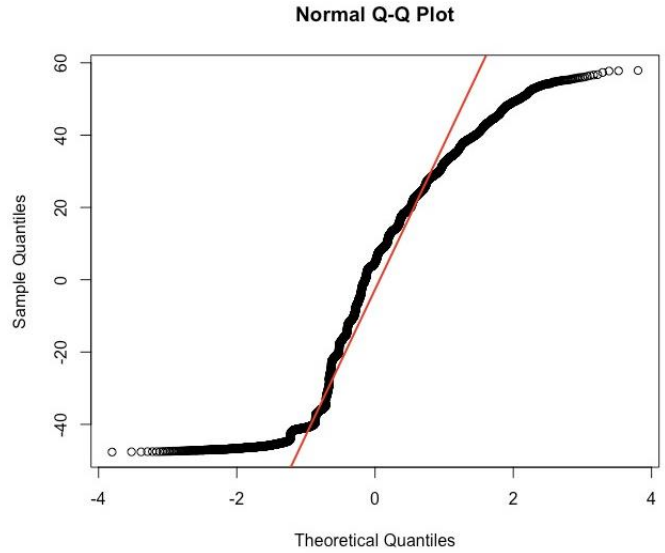

Figure 11: Normal Q-Q plot of residuals

The residuals of 5000+ samples do not seem normal (evident from the S-curve with huge amount of outliers). Furthermore, p-value of Shapiro-wilk test is calculated and is found to be smaller than $\alpha = 0.05$ which means that the residuals are not normally distributed. This test, in case of this data set is not completely reliable to its limitations of computing only 5000 residuals at a time, hence we go ahead with Anderson-Darling test and found similar results from that as well. The results from these tests are as follows:

```
        Shapiro-Wilk normality test

data:  data$resid2[1:5000]
W = 0.93424, p-value < 2.2e-16
```

```
        Anderson-Darling normality test

data:  data$resid1
A = 34.627, p-value < 2.2e-16
```

Figure 12: Normality tests

To conclude, the normality assumption has been violated.

2. Assumption of Homoscedasticity:
   Residuals Vs Fitted values plot has been plotted below which shows a "funnel" pattern indicating that this graph might not be a good fit for the data.

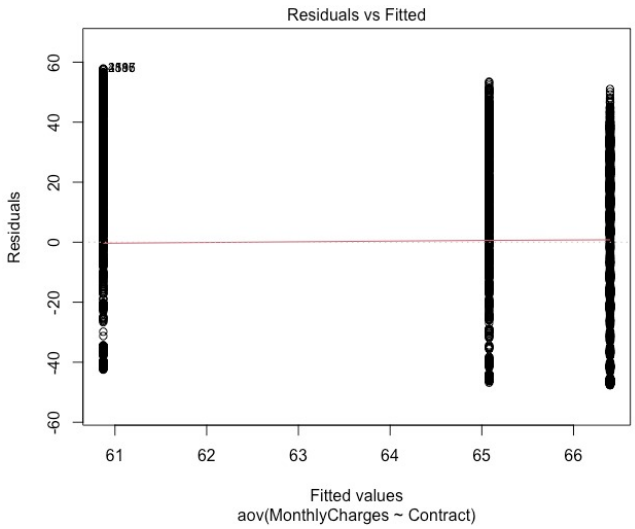

Figure 13: Residuals Vs Fitted values.

```
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    2  195.15 < 2.2e-16 ***
      7029
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 14: Levene's Test for equal variances

From the output of Levene's test (p-value < $\alpha = 0.05$) and nature of Residuals Vs Fitted values plot, it is concluded that the assumption of Homoscedasticity has been violated.

For us to perform One-way ANOVA, certain assumptions must be met -

- IID: we have k random samples collected from their respective populations.
- Independent Samples: the k samples are collected independently of one another.
- Normal Pop.: the k populations are all normally distributed.
- Homogeneity of Var.: the k populations possess the property of homoscedasticity (they all have the same population variance)

From the Q-Q plot we conclude that the datapoints are not normal. Furthermore, from the Levene's test we conclude that the data violates the assumption of homoscedasticity. Since both these important assumptions have been violated, One-way ANOVA and Welch's ANOVA tests cannot be conducted for this analysis. This is the reason Kruskal-Wallis test has been selected for ANOVA.

Kruskal-Wallis ANOVA:

Null hypothesis (H0): There is no significant difference in the median ranks of the populations from which the groups were sampled.

Alternative hypothesis (Ha): At least one group differs significantly from the others in terms of median ranks.

With all the assumptions of Kruskal-Wallis ANOVA test confirmed, test is conducted with following output:

```
        Kruskal-Wallis rank sum test

data:  MonthlyCharges by Contract
Kruskal-Wallis chi-squared = 19.997, df = 2, p-value = 4.546e-05
```

Figure 15: Kruskal-Wallis rank sum test output

With a moderate significance level of 0.05, it can be concluded that the null hypothesis can be rejected confirming that at least one group differs significantly from the others in terms of median ranks.

Now that we know atleast one median rank is different, it is suitable to conduct a post hoc multiple comparison procedure i.e., Dunn's test to compare all pairs of medians and report their p-values which will be compared with a significance level of $\alpha = 0.05$.

```
                              Comparison of x by group
                                   (No adjustment)
Col Mean-|
Row Mean |   Month-to   One year
---------+----------------------
One year |   0.556494
         |     0.2889
         |
Two year |   4.408700   3.128449
         |    0.0000*    0.0009*

alpha = 0.05
Reject Ho if p <= alpha/2
```

Figure 16: Dunn's test output

The asterisk next to the Two year - One year and Two year – Month-to-month comparison suggests that there is a significant difference between the medians of these two groups at the 95% level of significance. This is also evident from the parallel boxplots in Figure 9.

4. Conclusion:

After performing both the statistical analyses, we have summarized the following findings:

Analysis #1: The data set for Internet Service vs Monthly Charges violates the homoscedasticity assumption. Furthermore, the normal Q-Q plot shows the data to be fairly normal, however since the sample size of this data set is huge, it is safe to assume that the Welch's ANOVA test is robust to the violation of the normality assumption. Hence, Welch's ANOVA test is conducted, and it is found that there exists atleast one population mean which is significantly different than others. Furthermore, Pairwise t-test is conducted for multiple comparisons and concluded that the mean of Fibre optic is significantly different from other two types of internet services opted/not opted by the customers.

Analysis #2: The data set for Contract vs Monthly Charges violates both the normality and homoscedasticity assumptions. Hence, the Kruskal-Wallis test is conducted, and it is found that there is atleast one pair of group median ranks significantly different from others. Upon conducting Dunn's post hoc comparison test, it is found that Two year - One year and Two year – Month-to-month groups have significantly different median ranks.