

Chat Bot Human Evaluation Example

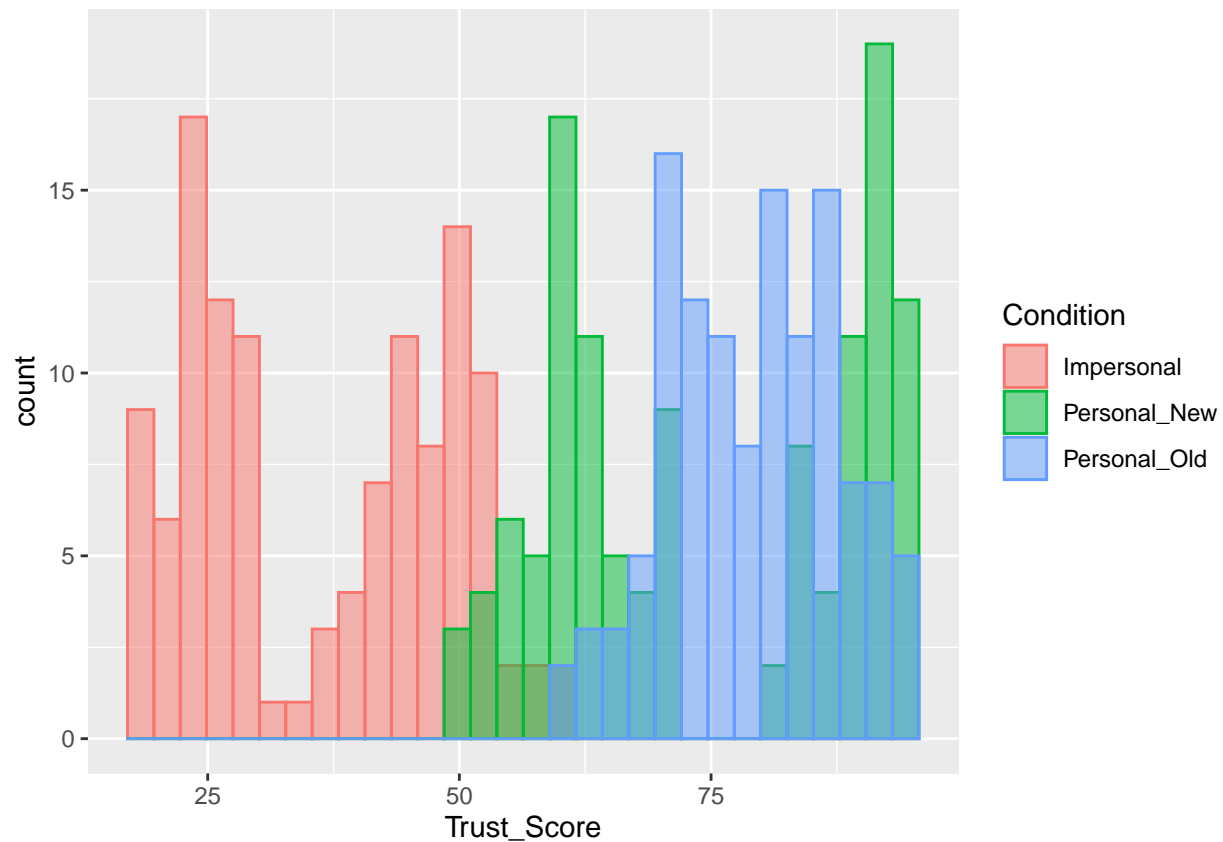
Toy Data

```
df <- read.csv("toy_data.csv", header=TRUE, stringsAsFactors=FALSE)
```

Overview Plot

```
library(ggplot2)
ggplot(df, aes(x=Trust_Score, fill=Condition, color=Condition)) +
  geom_histogram(position="identity", alpha=0.5)

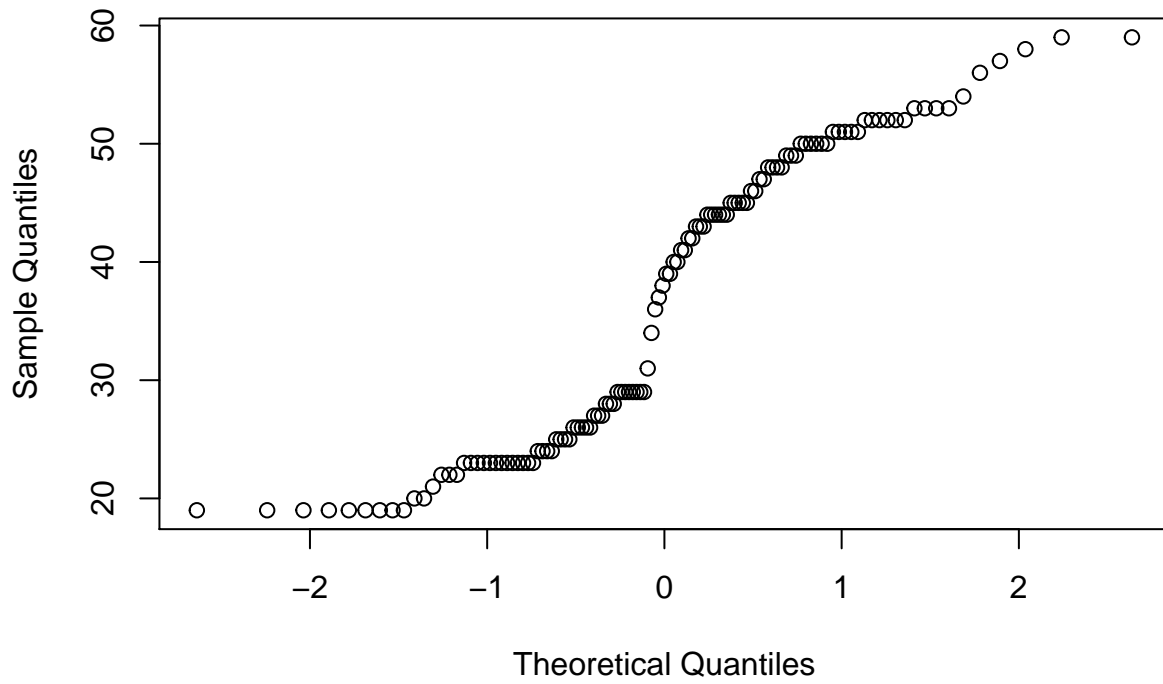
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Q-Q Plot

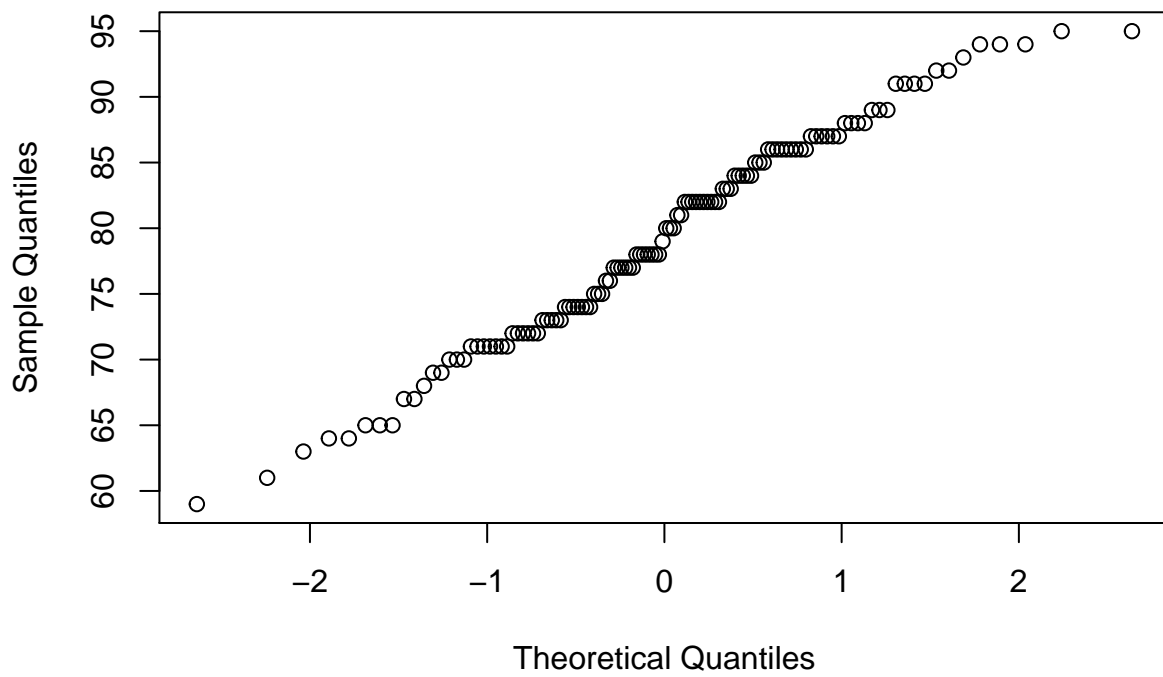
```
ImpersonalData <- subset(df, Condition=='Impersonal')
PersonalOldData <- subset(df, Condition=='Personal_Old')
PersonalNewData <- subset(df, Condition=='Personal_New')
qqnorm(ImpersonalData$Trust_Score)
```

Normal Q-Q Plot



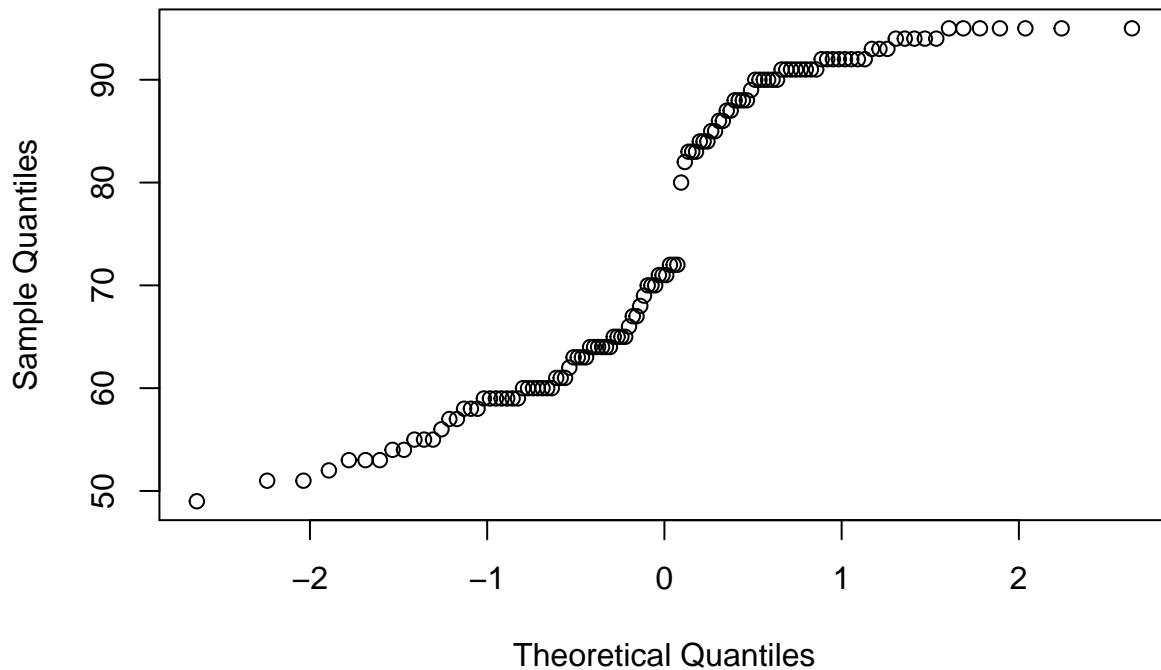
```
qqnorm(PersonalOldData$Trust_Score)
```

Normal Q-Q Plot



```
qqnorm(PersonalNewData$Trust_Score)
```

Normal Q-Q Plot



Friedmann Test

```
friedman.test(Trust_Score ~ Condition | Participant_Id, data = df)
```

```
##  
##  Friedman rank sum test  
##  
## data:  Trust_Score and Condition and Participant_Id  
## Friedman chi-squared = 182.33, df = 2, p-value < 2.2e-16
```

Post Hoc Test

```
library(PMCMRplus)  
frdAllPairsNemenyiTest(Trust_Score ~ Condition | Participant_Id, data = df)
```

```
##  
##  Pairwise comparisons using Nemenyi-Wilcoxon-Wilcox all-pairs test for a two-way balanced complete b  
## data: Trust_Score and Condition and Participant_Id  
##  
##           Impersonal Personal_New  
## Personal_New 2.8e-14      -  
## Personal_Old < 2e-16    0.19  
##  
## P value adjustment method: single-step
```