DSC530FinalSMITH

Champlin Smith

2025-03-02

**DSC530 Final Project**

# Initial basic model with every feature

**library**(ggplot2) **library**(dplyr)

##

## Attaching package: ’dplyr’

## The following objects are masked from ’package:stats’:

##

## filter, lag

## The following objects are masked from ’package:base’:

##

## intersect, setdiff, setequal, union

**library**(car)

## Loading required package: carData

##

## Attaching package: ’car’

## The following object is masked from ’package:dplyr’:

##

## recode

|  |
| --- |
| data <- **read.csv**("C:**\\**Users**\\**champ**\\**OneDrive**\\**Documents**\\**songs.csv")  model <- **lm**(track\_popularity **~** GenreInt **+** danceability **+** energy **+** key **+** loudness **+ summary**(model) |

speechiness **+** acousti

##

## Call:

## lm(formula = track\_popularity ~ GenreInt + danceability + energy +

## key + loudness + speechiness + acousticness + liveness +

## valence + tempo + duration\_ms, data = data)

##

## Residuals:

## Min 1Q Median 3Q Max

## -0.31096 -0.14098 -0.04386 0.09023 0.77031

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 2.841e-01 7.312e-02 3.886 0.000111 \*\*\*

## GenreInt -1.033e-02 1.759e-03 -5.872 6.55e-09 \*\*\* ## danceability 9.422e-02 4.440e-02 2.122 0.034183 \*

## energy -2.578e-01 6.436e-02 -4.006 6.80e-05 \*\*\*

|  |  |
| --- | --- |
| ## key 3.532e-02 2.064e-02 | 1.711 0.087498 . |
| ## loudness 1.842e-01 8.972e-02 | 2.053 0.040385 \* |
| ## speechiness 6.334e-02 5.638e-02 | 1.124 0.261573 |
| ## acousticness 1.220e-02 3.467e-02 | 0.352 0.724952 |
| ## liveness 7.129e-03 4.972e-02 | 0.143 0.886023 |

## valence -2.474e-02 3.535e-02 -0.700 0.484158

## tempo 2.141e-02 3.443e-02 0.622 0.534179

## duration\_ms -2.568e-07 1.553e-07 -1.654 0.098551 .

## ---

## Signif. codes: 0 ’\*\*\*’ 0.001 ’\*\*’ 0.01 ’\*’ 0.05 ’.’ 0.1 ’ ’ 1

##

## Residual standard error: 0.1866 on 728 degrees of freedom

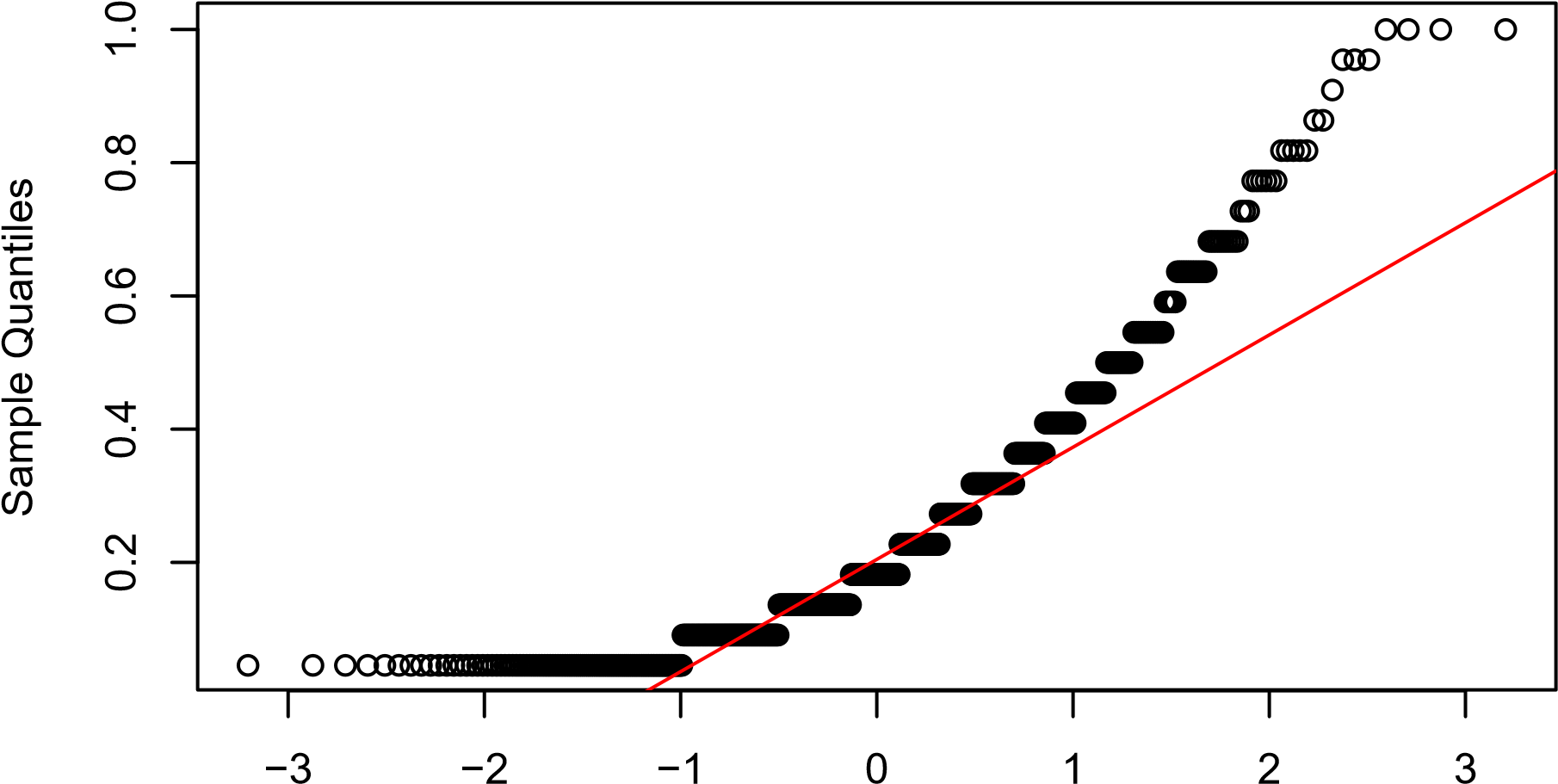
## Multiple R-squared: 0.09598, Adjusted R-squared: 0.08232

## F-statistic: 7.027 on 11 and 728 DF, p-value: 2.194e-11

# QQ-Plot

*# Q-Q Plot for Track Popularity* **qqnorm**(data**$**track\_popularity, main="Q-Q Plot of Track Popularity") **qqline**(data**$**track\_popularity, col="red") *# Adds a reference line*

## **Q−Q Plot of Track Popularity**

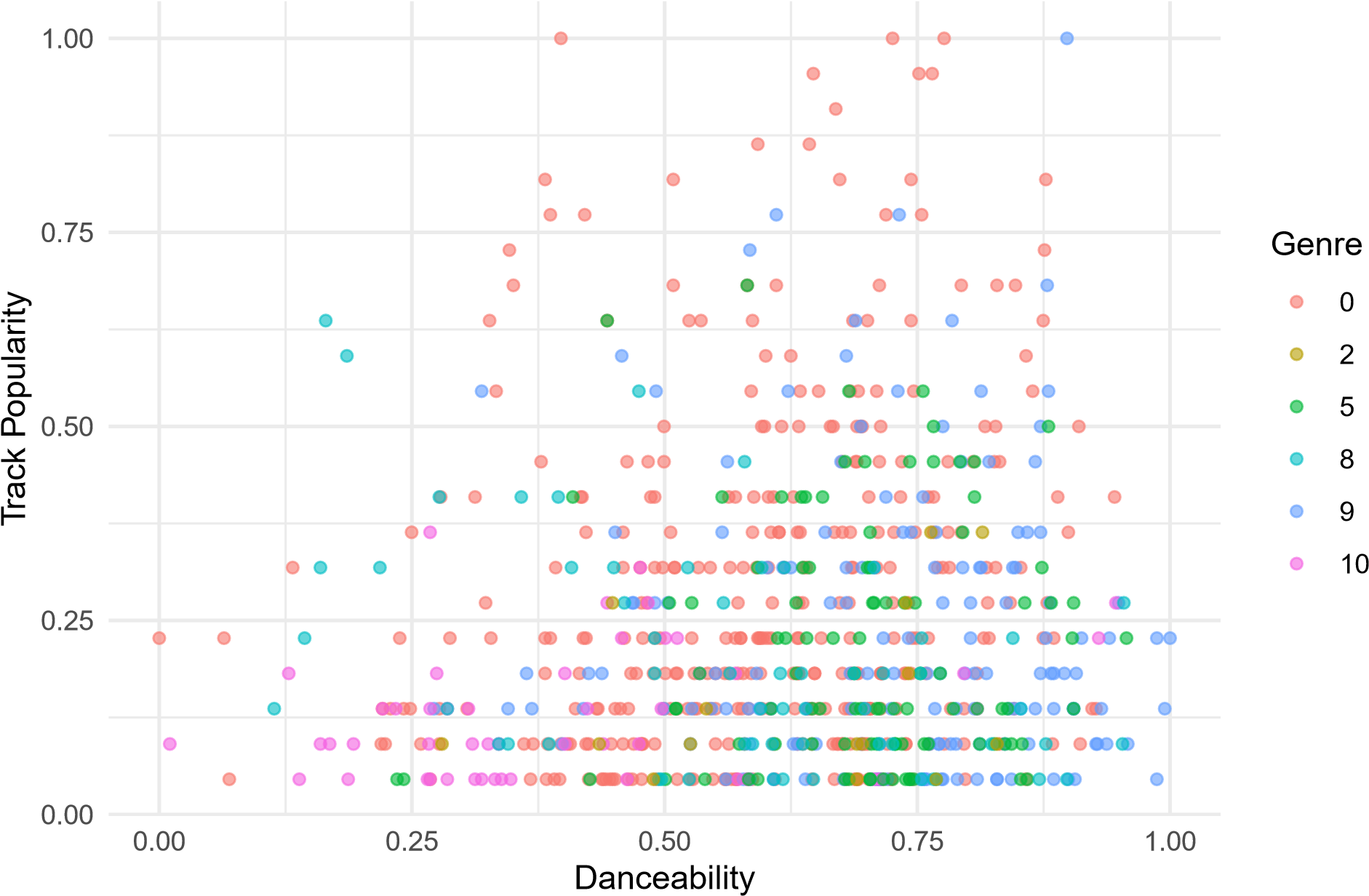


Theoretical Quantiles

#Scatter plot

|  |
| --- |
| **library**(ggplot2)  *# Scatter Plot of Popularity vs. Danceability with Genre as Color* **ggplot**(data, **aes**(x = danceability, y = track\_popularity, color = **as.factor**(GenreInt))) **+**  **geom\_point**(alpha = 0.6) **+** *# Scatter points with some transparency* **labs**(title = "Track Popularity vs. Danceability", x = "Danceability", y = "Track Popularity", color = "Genre") **+** *# Legend title*  **theme\_minimal**() |

### Track Popularity vs. Danceability



# Histograms

numeric\_vars <- **names**(data)[**sapply**(data, is.numeric)]

**par**(mfrow = **c**(3, 4)) **for** (var **in** numeric\_vars) {

**hist**(data[[var]], main = **paste**("Histogram of", var), xlab = var, col = "lightblue", border = "black")

}

**Histogram of X Histogram of track\_popul Histogram of GenreIn Histogram of danceabil**

Frequency

0

80

Frequency

0

200

Frequency

0

300

Frequency

0

150

1. 400 800 0.0 0.4 0.8 0 4 8 0.0 0.4 0.8

X track\_popularity GenreInt danceability

**Histogram of energy Histogram of key Histogram of loudness Histogram of mode**

Frequency

0

150

Frequency

0

150

Frequency

0

300

Frequency

0

400

0.0 0.4 0.8 0.0 0.4 0.8 0.0 0.4 0.8 0.0 0.4 0.8

energy key loudness mode

**Histogram of speechine Histogram of acousticne Histogram of liveness Histogram of valence**

Frequency

0

400

Frequency

0

250

Frequency

0

300

Frequency

0

80

|  |  |  |  |
| --- | --- | --- | --- |
| 0.0 0.4 0.8  speechiness | 0.0 0.4 0.8  acousticness | 0.0 0.4 0.8  liveness | 0.0 0.4 0.8  valence |
| **par**(mfrow = **c**(1, 1)) |  |  |  |

**Histogram of tempo Histogram of duration\_m**

Frequency

0

150

Frequency

0

250

0.0 0.4 0.8 1e+05 4e+05

tempo duration\_ms

# Descriptive statistics, skewness, kurtosis

|  |
| --- |
| **library**(moments)  *# Function to compute descriptive statistics* compute\_stats <- **function**(var) { var\_data <- data[[var]]  mean\_val <- **mean**(var\_data, na.rm = TRUE) mode\_val <- **as.numeric**(**names**(**sort**(**table**(var\_data), decreasing = TRUE)[1])) sd\_val <- **sd**(var\_data, na.rm = TRUE) iqr\_val <- **IQR**(var\_data, na.rm = TRUE) skew\_val <- **skewness**(var\_data, na.rm = TRUE) *# Skewness for tail behavior* kurt\_val <- **kurtosis**(var\_data, na.rm = TRUE) *# Kurtosis for peak and tail thickness*  **return**(**data.frame**(Variable = var, Mean = mean\_val, Mode = mode\_val,  SD = sd\_val, IQR = iqr\_val, Skewness = skew\_val, Kurtosis = kurt\_val))  }  *# Apply to all numeric variables*  numeric\_vars <- **names**(data)[**sapply**(data, is.numeric)] stats\_df <- **do.call**(rbind, **lapply**(numeric\_vars, compute\_stats))  *# Print summary statistics* **print**(stats\_df) |

*# Get most frequent value*

## Variable Mean Mode SD IQR

## 1 X 4.394257e+02 0.000000e+00 2.614860e+02 4.575000e+02

## 2 track\_popularity 2.426904e-01 4.545455e-02 1.947937e-01 2.272727e-01

|  |  |
| --- | --- |
| ## 3 | GenreInt 3.947297e+00 0.000000e+00 4.123753e+00 9.000000e+00 |
| ## 4 | danceability 6.238226e-01 4.901961e-01 1.835794e-01 2.503268e-01 |
| ## 5 | energy 6.380446e-01 5.675502e-01 1.789285e-01 2.438566e-01 |
| ## 6 | key 4.883292e-01 9.090909e-02 3.334232e-01 6.363636e-01 |
| ## 7 | loudness 7.266619e-01 7.311468e-01 1.129410e-01 1.194641e-01 |
| ## 8 | mode 5.729730e-01 1.000000e+00 4.949808e-01 1.000000e+00 |
| ## 9 | speechiness 1.190647e-01 9.730899e-03 1.340607e-01 1.505586e-01 |
| ## 10 | acousticness 2.272270e-01 1.502358e-01 2.374978e-01 2.891530e-01 |
| ## 11 | liveness 1.594110e-01 9.446736e-02 1.406967e-01 1.094891e-01 |
| ## 12 | valence 5.135073e-01 6.232289e-01 2.364009e-01 3.692572e-01 |
| ## 13 | tempo 4.246232e-01 4.592441e-01 2.095823e-01 3.193064e-01 |
| ## 14 | duration\_ms 2.098158e+05 1.721960e+05 4.645584e+04 5.192350e+04 |
| ## | Skewness Kurtosis |
| ## 1 | 0.04420222 1.789432 |
| ## 2 | 1.37846797 4.808525 |
| ## 3 | 0.25774423 1.280565 |

## 4 -0.57883038 3.059934

## 5 -0.56094743 3.082574

## 6 -0.02578939 1.648289

## 7 -1.77261995 9.399089

## 8 -0.29505113 1.087055

## 9 1.66675425 6.200244

## 10 1.35883395 4.124408

## 11 2.64722842 12.248376

## 12 0.02443950 2.139053

## 13 0.45435935 2.348259

## 14 1.40849058 7.892719

# PDF Comparisons

|  |
| --- |
| genre\_x <- data **%>% filter**(GenreInt **==** 2) other\_genres <- data **%>% filter**(GenreInt **!=** 2)  compute\_pmf <- **function**(df, variable) {  pmf <- df **%>%**  **count**(.data[[variable]]) **%>% mutate**(Probability = n **/ sum**(n))  **return**(pmf)  }  pmf\_genre\_x <- **compute\_pmf**(genre\_x, "track\_popularity") pmf\_other\_genres <- **compute\_pmf**(other\_genres, "track\_popularity")  pmf\_genre\_x**$**Scenario <- "EDM" pmf\_other\_genres**$**Scenario <- "Other Genres" pmf\_combined <- **bind\_rows**(pmf\_genre\_x, pmf\_other\_genres) **ggplot**(pmf\_combined, **aes**(x = track\_popularity, y = Probability, fill = Scenario)) **+** |

**geom\_bar**(stat = "identity", position = "dodge", alpha = 0.7) **+ labs**(title = "PMF of Track Popularity by Genre", x = "Track Popularity", y = "Probability") **+**

**theme\_minimal**()

## PMF of Track Popularity by Genre

0.0

0.1

0.2

0.3

0.4

0.00

0.25

0.50

0.75

1.00

Track Popularity

Probability

Scenario

EDM

Other Genres

# CDF

|  |
| --- |
| **ggplot**(data, **aes**(x = track\_popularity)) **+**  **stat\_ecdf**(geom = "step", color = "blue", size = 1) **+**  **labs**(title = "Cumulative Distribution Function (CDF) of Track Popularity", x = "Track Popularity", y = "Cumulative Probability") **+**  **theme\_minimal**() |

## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.

## i Please use ‘linewidth‘ instead.

## This warning is displayed once every 8 hours.

## Call ‘lifecycle::last\_lifecycle\_warnings()‘ to see where this warning was ## generated.

## Cumulative Distribution Function (CDF) of Track Popularity

0.00

0.25

0.50

0.75

1.00

0.00

0.25

0.50

0.75

1.00

Track Popularity

Cumulative Probability

# Analytic distrbution - histogram vs normal distribution

|  |
| --- |
| pop\_mean <- **mean**(data**$**track\_popularity, na.rm = TRUE) pop\_sd <- **sd**(data**$**track\_popularity, na.rm = TRUE)  **ggplot**(data, **aes**(x = track\_popularity)) **+ geom\_histogram**(**aes**(y = ..density..), bins = 30, fill = "lightblue", color = "black", **stat\_function**(fun = dnorm, args = **list**(mean = pop\_mean, sd = pop\_sd), color = "red", size = 1) **+**  **labs**(title = "Histogram of Track Popularity with Normal Distribution", x = "Track Popularity", y = "Density") **+**  **theme\_minimal**() |

alpha = 0.7) **+**

## Warning: The dot-dot notation (‘..density..‘) was deprecated in ggplot2 3.4.0.

## i Please use ‘after\_stat(density)‘ instead.

## This warning is displayed once every 8 hours.

## Call ‘lifecycle::last\_lifecycle\_warnings()‘ to see where this warning was ## generated.

## Histogram of Track Popularity with Normal Distribution

0

1

2

3

4

5

0.00

0.25

0.50

0.75

1.00

Track Popularity

Density

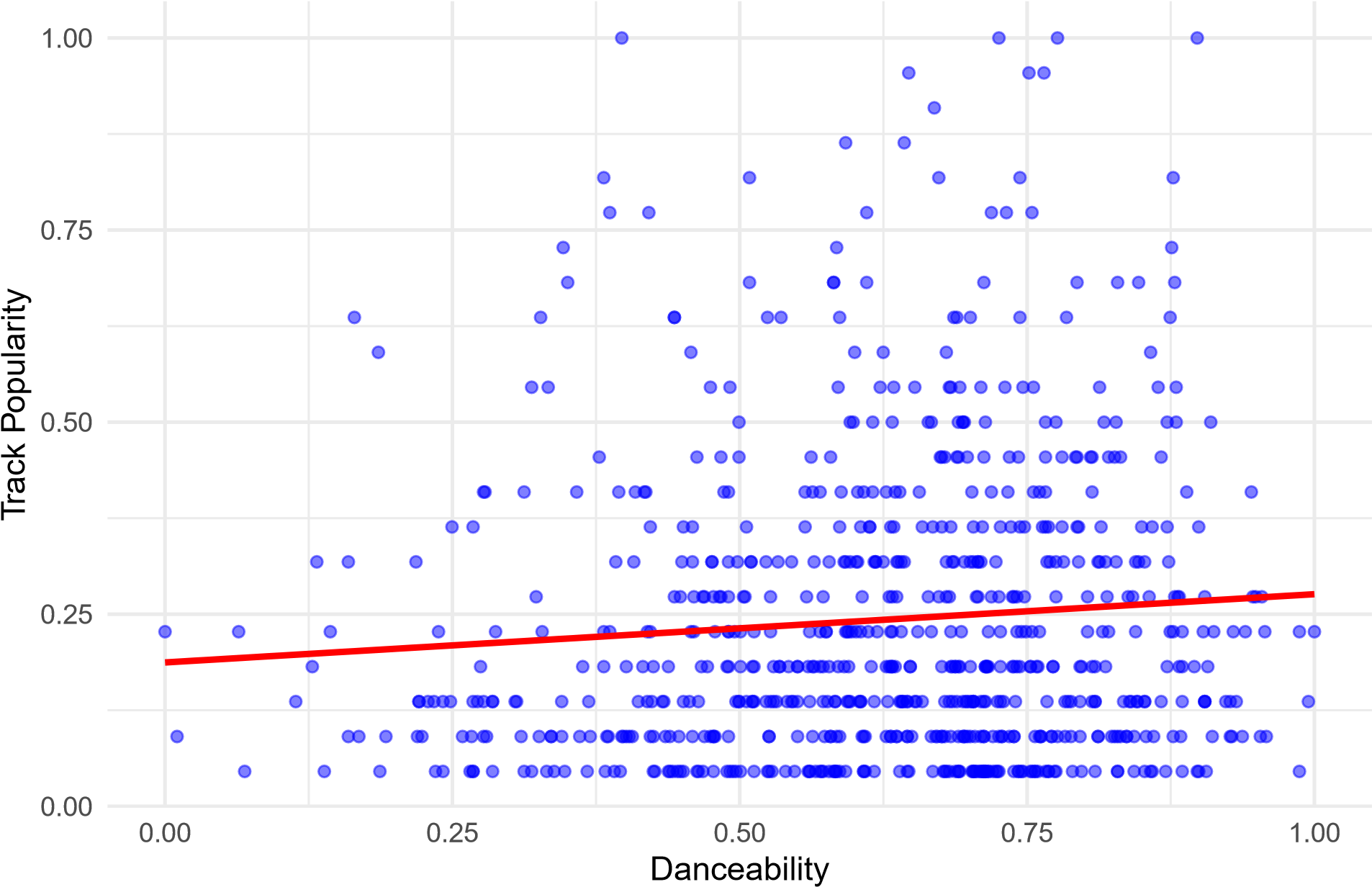
# The distribution of popularity in the dataset is not normal as seen in the above analytic. It is skewed positively and has a kurtosis nearr 5. It has a very heavy tail initially

# Scatter - Danceability

|  |
| --- |
| **ggplot**(data, **aes**(x = danceability, y = track\_popularity)) **+**  **geom\_point**(alpha = 0.5, color = "blue") **+**  **geom\_smooth**(method = "lm", color = "red", se = FALSE) **+ labs**(title = "Scatter Plot: Danceability vs. Track Popularity", x = "Danceability", y = "Track Popularity") **+**  **theme\_minimal**() |

## ‘geom\_smooth()‘ using formula = ’y ~ x’

## Scatter Plot: Danceability vs. Track Popularity

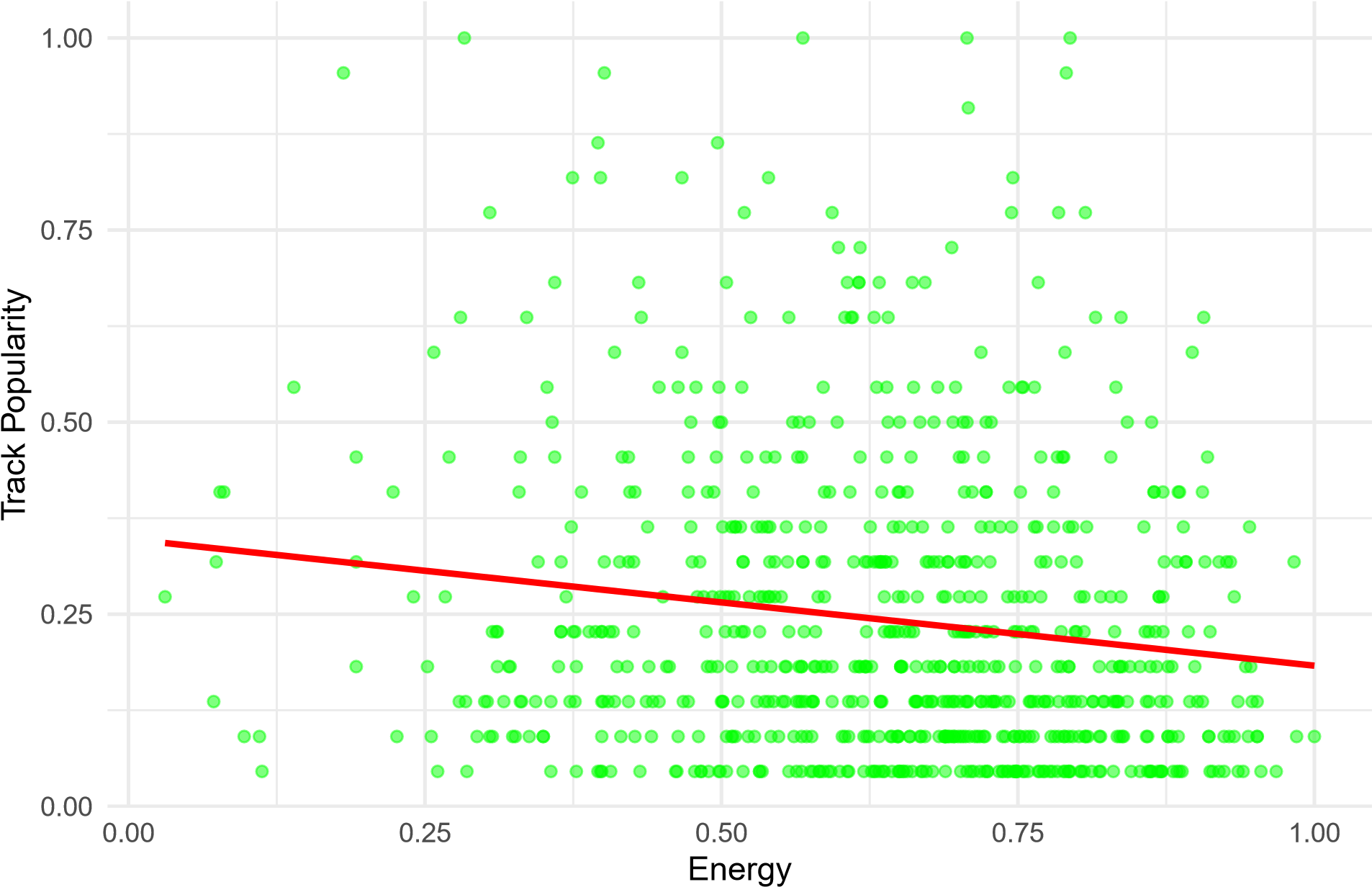


# Scatter - Energy

|  |
| --- |
| **ggplot**(data, **aes**(x = energy, y = track\_popularity)) **+**  **geom\_point**(alpha = 0.5, color = "green") **+ geom\_smooth**(method = "lm", color = "red", se = FALSE) **+ labs**(title = "Scatter Plot: Energy vs. Track Popularity", x = "Energy", y = "Track Popularity") **+**  **theme\_minimal**() |

## ‘geom\_smooth()‘ using formula = ’y ~ x’

## Scatter Plot: Energy vs. Track Popularity



# Correlation Matrix

|  |
| --- |
| cor\_matrix <- **cor**(data[, **c**("danceability", "energy", "track\_popularity")], use = **print**("Correlation") |

"complete.obs", method

## [1] "Correlation"

**print**(cor\_matrix)

## danceability energy track\_popularity

## danceability 1.00000000 0.03688744 0.08370932

## energy 0.03688744 1.00000000 -0.15127816

## track\_popularity 0.08370932 -0.15127816 1.00000000

|  |
| --- |
| cov\_matrix <- **cov**(data[, **c**("danceability", "energy", "track\_popularity")], use = **print**("Covariance") |

"complete.obs")

## [1] "Covariance"

**print**(cov\_matrix)

## danceability energy track\_popularity

## danceability 0.033701396 0.001211663 0.002993454

## energy 0.001211663 0.032015406 -0.005272670

## track\_popularity 0.002993454 -0.005272670 0.037944574

**The better to understand measure to determine the strength and direction of the relationship between varialbes in Pearsons Correlation, since it provides a set scale from -1 to 1 instead of raw empirical numbers. With this understanding, I do not see any multicollinearity issues as no variables are highly correlated with each other. The highest correlation is -.14, and this is well within bounds to be included in modeling as a feature.**

**Hypothsis Testing. Hypothesis - Energy has no affect on song popularity.**

model <- **lm**(track\_popularity **~** GenreInt **+** danceability **+** energy **+** key **+** loudness **+** speechiness **+** acousti **summary**(model)

##

## Call:

## lm(formula = track\_popularity ~ GenreInt + danceability + energy +

## key + loudness + speechiness + acousticness + liveness +

## valence + tempo + duration\_ms, data = data)

##

## Residuals:

## Min 1Q Median 3Q Max

## -0.31096 -0.14098 -0.04386 0.09023 0.77031

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 2.841e-01 7.312e-02 3.886 0.000111 \*\*\*

## GenreInt -1.033e-02 1.759e-03 -5.872 6.55e-09 \*\*\* ## danceability 9.422e-02 4.440e-02 2.122 0.034183 \*

## energy -2.578e-01 6.436e-02 -4.006 6.80e-05 \*\*\* ## key 3.532e-02 2.064e-02 1.711 0.087498 .

## loudness 1.842e-01 8.972e-02 2.053 0.040385 \*

## speechiness 6.334e-02 5.638e-02 1.124 0.261573

## acousticness 1.220e-02 3.467e-02 0.352 0.724952

## liveness 7.129e-03 4.972e-02 0.143 0.886023

## valence -2.474e-02 3.535e-02 -0.700 0.484158

## tempo 2.141e-02 3.443e-02 0.622 0.534179 ## duration\_ms -2.568e-07 1.553e-07 -1.654 0.098551 . ## ---

## Signif. codes: 0 ’\*\*\*’ 0.001 ’\*\*’ 0.01 ’\*’ 0.05 ’.’ 0.1 ’ ’ 1

##

## Residual standard error: 0.1866 on 728 degrees of freedom

## Multiple R-squared: 0.09598, Adjusted R-squared: 0.08232

## F-statistic: 7.027 on 11 and 728 DF, p-value: 2.194e-11

**Energy is statistically significant, and thus we reject the null hypothesis. Energy has a statistically significant affect on song popularity.**

# Step-wise feature selection

model <- **lm**(track\_popularity **~** GenreInt **+** danceability **+** energy **+** key **+** loudness **+** duration\_ms, data) **summary**(model)

##

## Call:

## lm(formula = track\_popularity ~ GenreInt + danceability + energy +

## key + loudness + duration\_ms, data = data)

##

## Residuals:

## Min 1Q Median 3Q Max

## -0.30296 -0.13848 -0.04645 0.09141 0.77835

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 3.142e-01 6.409e-02 4.903 1.16e-06 \*\*\*

## GenreInt -9.713e-03 1.703e-03 -5.702 1.71e-08 \*\*\* ## danceability 8.714e-02 3.904e-02 2.232 0.0259 \*

## energy -2.749e-01 5.532e-02 -4.969 8.37e-07 \*\*\* ## key 3.667e-02 2.056e-02 1.784 0.0749 .

## loudness 1.798e-01 8.891e-02 2.022 0.0435 \* ## duration\_ms -2.893e-07 1.529e-07 -1.892 0.0589 .

## ---

## Signif. codes: 0 ’\*\*\*’ 0.001 ’\*\*’ 0.01 ’\*’ 0.05 ’.’ 0.1 ’ ’ 1

##

## Residual standard error: 0.1863 on 733 degrees of freedom

## Multiple R-squared: 0.09293, Adjusted R-squared: 0.0855

## F-statistic: 12.52 on 6 and 733 DF, p-value: 1.845e-13