BRITISH AIRWAYS

ANALYSIS OF WEB SCRAPPED BRITISH AIRLINE REVIEWS

```
library(rvest)
library(janitor)
library(tidyverse)
library(data.table)
library(RSelenium)
library(rlist)
library(utf8)
library(VIM)
```

ANALYSIS: Data was extracted from website https://www.airlinequality.com/airline-reviews/british-airways

Reading the data

```
category_reviews <- read_csv("category_reviews.csv")
overal_reviews <- read_csv("overal_reviews.csv")
```

Remove the id column from category_reviews;

```
category_reviews <- category_reviews %>%
    select(-id)
```

table_id in category_reviews represent the id of each review-number of unique table_id;

```
range(category_reviews$table_id)
```

```
[1]    1 3000
```

table_id is recycled from 1 to 3000 representing the first review to the 3000th review.

overal_reviews has 3350 reviews but category review has 3000 reviews

1. **For the overal_rating**;

1.1 **Review dates**

```
overal_reviews_1 <- overal_reviews %>%
    mutate(review_dates = dmy(review_dates))
range(overal_reviews_1$review_dates)
```

```
[1] "2015-03-04" "2025-05-18"
```

The reviews were from 04/03/2015 to 18/05/2025.

1.2 **Trip Verification**

```
knitr::kable(overal_reviews_1 %>%
    count(trip_occurrence), caption = "TRAVELLERS", longtable = TRUE,
    digits = 2, format.args = list(big.mark = ",", scientific = FALSE),
    "latex") %>%
    kableExtra::column_spec(1, border_left = T) %>%
    kableExtra::column_spec(2, border_right = T) %>%
    kableExtra::kable_styling(latex_options = "HOLD_position",
        "repeat_header")
```

Table 1: TRAVELLERS

trip_occurrence	n
Not Verified	1,252
Verified	2,098

It was verified that out of 3350 reviews 2098 were of verified individuals who actually flew the airline.

1.3 Verified 2098 travellers

```
overall_verified <- overal_reviews_1 %>%
    filter(trip_occurrence == "Verified")
overall_verified$reviewer_name <- str_to_upper(overall_verified$reviewer_name)  ### names to u
overall_verified$reviewer_name <- str_squish(overall_verified$reviewer_name)  ### remove extra
```

1.3.1 Unique travellers

```
frequent_travellers <- overall_verified %>%
    count(reviewer_name) %>%
    arrange(desc(n)) %>%
    filter(n > 1)
```

Number of individuals that had travelled more than once

```
nrow(frequent_travellers)
```

```
[1] 173
```

Top most travellers

```
knitr::kable(frequent_travellers, caption = "REPEAT TRAVELLERS",
    longtable = TRUE, digits = 2, format.args = list(big.mark = ",",
        scientific = FALSE), "latex") %>%
    kableExtra::column_spec(1, border_left = T) %>%
    kableExtra::column_spec(2, border_right = T) %>%
    kableExtra::kable_styling(latex_options = "HOLD_position",
        "repeat_header")
```

Table 2: REPEAT TRAVELLERS

reviewer_name	n
E SMYTH	34
DAVID ELLIS	28
CLIVE DRAKE	13
CHRISTOPHER NEEP	10
JOHN ROLFE	10
C DOWN	9
DAVID TAYLOR	9
J MEARES	8
R VINES	8
RICHARD HODGES	8
MIKE PALMER	7
A WONG	6
ALISTAIR BAKER	6
C BARTON	6
DEREK NORTHCUTT	6
ALLY WHARTON	5
ANGELO MENEZES	5
D GOLD	5
JOHN PRESCOTT	5
JONATHAN RODDEN	5
M EDWARDS	5
MARK ELLWOOD	5
ROBERT WATSON	5
VINCENT BORLAUG	5
ALAN THOMPSON	4
ALWALEED ALTHANI	4
ANDY MAGOWAN	4
C FORDHAM	4
D WEBB	4
G GRAHAM	4
GUSTAVO SIRNA BARBOSA	4
IAN ROBINSON	4
IAN SINCLAIR	4
KATHLEEN KIRBY	4

NEIL JERAM	4
PAUL MERCER	4
R SCHRÖDER	4
RAJAN PARRIKAR	4
A AHMED	3
A DAWSON	3
A WHARTON	3
ALAN WAN	3
ALBERT WONG	3
B SAUNDERS	3
B STAUFFER	3
C CORDAN	3
C KAY	3
C PORTER	3
C RANKIN	3
CALEB LOWE	3
COLIN PAY	3
CRAIG CUTTS	3
E MICHAELS	3
HARRY ARONOWICZ	3
J FANG	3
J HUGO	3
KAH KAY AU	3
LUIS DE JESUS	3
M BEALE	3
M WILLIAMS	3
MAHMUD NOORMOHAMED	3
MICHAEL PAPALAMPROU	3
MICHAEL SCHADE	3
MIKE ANDRÉSEN	3
PETER POMERANZE	3
R GONZAGA	3
S ANDERSON	3
S GRAHAM	3
S PORTER	3
SARAH SHAILES	3
SIMON CHANNON	3
A LEWIS	2
A MALTAM	2
A NORTON	2
A PALOMO	2
A WARD	2
ALLAN GITTENS	2
AMANDA EDGAR	2
ANDERS PEDERSEN	2

ANDREW PYBUS	2
ATTILA TOTH	2
B MEARES	2
B STEWART	2
C ANDREWS	2
C BEALE	2
C BOWEN	2
C DEAN	2
C DREW	2
C HOFFMANN	2
C LANE	2
C LEARE	2
C STAINER	2
C STRATTON	2
CHRIS WALSH	2
CHRISTOPHER RAINBOW	2
CHUN SING POON	2
COLIN BARRY	2
D GORDON	2
D LEWIS	2
E CARMERE	2
E LANDEN	2
EELCO VAN DEN HEUVEL	2
G JONES	2
G LEANE	2
GIOVANNI GIORGIS	2
GLENN BIFFEN	2
GLENN TAYLOR-BIFFEN	2
H BURTON	2
H JACKSON	2
H LIND	2
H MILLER	2
H NEALE	2
ISHAN PAI	2
J FORLEN	2
J PEARCE	2
JANEKS VOLKOVS	2
JOHN BARRY	2
K HAYMES	2
K ROBINSON	2
KEIRAN COULTON	2
KEN HOWIE	2
L HARPER	2
L IRVING	2
L RENNIE	2

L TRAN	2
M CAMERE	2
M DAVIDSON	2
M HART	2
M IRVING	2
M KEMP	2
M OWEN	2
MANUEL VIEIRA	2
MARK MCCULLOUGH	2
MATEUSZ WALTER	2
N ANDERSON	2
N CARTER	2
NUNO LUZ	2
OWEN BERKELEY-HILL	2
P ANDREWS	2
P GARVEY	2
P GOUGH	2
P MARTEN	2
P TYLER	2
PAUL RENSHAW	2
PETER COSTELLO	2
R ANDERSON	2
R DAWSON	2
R HEALE	2
R MARTON	2
R NEALE	2
R SANYAL	2
RICHARD CALLIS	2
ROBERT DAVIS	2
ROHITH JAYAWARDENE	2
S BEALE	2
S JOHNSON	2
S KEANE	2
S MORTON	2
S SIMPSON	2
S TEUGET	2
S WARD	2
S WARDEN	2
SIMON FOWLER	2
STEFAN VETTER	2
STEVEN HODGSON	2
T LEANE	2
T MEARES	2
TONY BANWAIT	2
TONY MCLAUGHLIN	2

W ANDERSON	2
W COLE	2
W HEALE	2
Y CHAN	2

frequent travelers that used British Airways on 2

```
frequent_return_only <- frequent_travellers %>%
    filter(n < 3)
nrow(frequent_return_only)
```

```
[1] 102
```

Range of rating

```
range(overall_verified$overal_rating)
```

```
[1]  1 10
```

the lowest rating was 1 while the highest rating was 10

the rating is an ordered factor from 1 to 10

```
overall_verified$overal_rating <- factor(overall_verified$overal_rating,
    levels = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10), ordered = T)
```

```
knitr::kable(overall_verified %>%
    count(overal_rating) %>%
    mutate(percent = n/sum(n) * 100), caption = "NO OF TRAVELLERS PER RATING",
    longtable = TRUE, digits = 2, format.args = list(big.mark = ",",
        scientific = FALSE), "latex") %>%
    kableExtra::column_spec(1, border_left = T) %>%
    kableExtra::column_spec(3, border_right = T) %>%
    kableExtra::kable_styling(latex_options = "HOLD_position",
        "repeat_header")
```

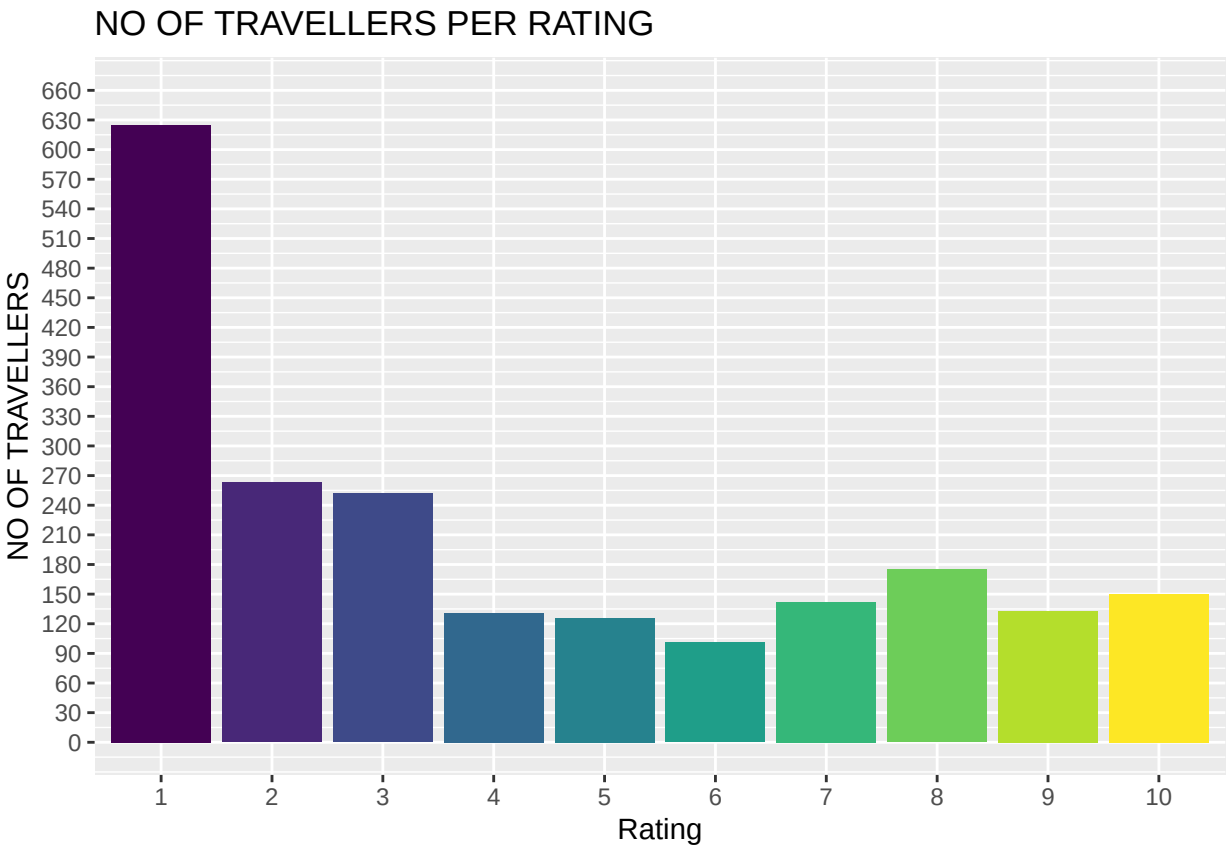
Table 3: NO OF TRAVELLERS PER RATING

overal_rating	n	percent

1	625	29.79
2	263	12.54
3	252	12.01
4	131	6.24
5	126	6.01
6	101	4.81
7	142	6.77
8	175	8.34
9	133	6.34
10	150	7.15

From the number of verified travellers, most gave the airline a very low overall rating of 1. Out of the 2098 travellers 1271 of them gave a rating of 4 and below.

```
overall_verified %>%
    count(overal_rating) %>%
    ggplot(aes(x = overal_rating,
        y = n, fill = overal_rating)) +
    geom_bar(stat = "identity",
        position = "dodge") +
    theme(legend.position = "none") +
    scale_y_continuous("NO OF TRAVELLERS",
        breaks = seq(0, 660, by = 30),
        limits = c(0, 660)) +
    labs(title = "NO OF TRAVELLERS PER RATING",
        x = "Rating")
```



Country

```
overall_verified$review_country <- str_to_upper(overall_verified$review_country)  ### names to
overall_verified$review_country <- str_squish(overall_verified$review_country)  ### remove ext
```

```
knitr::kable(overall_verified %>%
    count(review_country) %>%
    mutate(percent = n/sum(n) * 100) %>%
    arrange(desc(n)), caption = "COUNTRY OF TRAVELLERS", longtable = TRUE,
    digits = 2, format.args = list(big.mark = ",", scientific = FALSE),
    "latex") %>%
    kableExtra::column_spec(1, border_left = T) %>%
    kableExtra::column_spec(3, border_right = T) %>%
    kableExtra::kable_styling(latex_options = "HOLD_position",
        "repeat_header")
```

Table 4: COUNTRY OF TRAVELLERS

review_country	n	percent
UNITED KINGDOM	1,259	60.01

UNITED STATES	267	12.73
AUSTRALIA	70	3.34
CANADA	65	3.10
GERMANY	47	2.24
SOUTH AFRICA	27	1.29
SWITZERLAND	25	1.19
NETHERLANDS	21	1.00
SINGAPORE	19	0.91
CHINA	18	0.86
FRANCE	18	0.86
IRELAND	15	0.71
HONG KONG	14	0.67
SPAIN	14	0.67
SWEDEN	14	0.67
UNITED ARAB EMIRATES	14	0.67
BELGIUM	13	0.62
INDIA	13	0.62
GREECE	12	0.57
MALAYSIA	12	0.57
PORTUGAL	10	0.48
THAILAND	10	0.48
ITALY	9	0.43
POLAND	7	0.33
QATAR	7	0.33
DENMARK	6	0.29
GHANA	6	0.29
ICELAND	6	0.29
MEXICO	6	0.29
ARGENTINA	4	0.19
AUSTRIA	4	0.19
CYPRUS	4	0.19
CZECH REPUBLIC	4	0.19
JAPAN	4	0.19
NORWAY	4	0.19
SOUTH KOREA	4	0.19
BRAZIL	3	0.14
NEW ZEALAND	3	0.14
NIGERIA	3	0.14
ROMANIA	3	0.14
SAUDI ARABIA	3	0.14
HUNGARY	2	0.10
LEBANON	2	0.10
RUSSIAN FEDERATION	2	0.10
SLOVAKIA	2	0.10
TAIWAN	2	0.10

TURKEY	2	0.10
BERMUDA	1	0.05
BULGARIA	1	0.05
CAYMAN ISLANDS	1	0.05
CHILE	1	0.05
COSTA RICA	1	0.05
DOMINICAN REPUBLIC	1	0.05
ECUADOR	1	0.05
EGYPT	1	0.05
INDONESIA	1	0.05
ISRAEL	1	0.05
KUWAIT	1	0.05
LAOS	1	0.05
MOROCCO	1	0.05
PANAMA	1	0.05
PHILIPPINES	1	0.05
SAINT KITTS AND NEVIS	1	0.05
SENEGAL	1	0.05
UKRAINE	1	0.05
VIETNAM	1	0.05

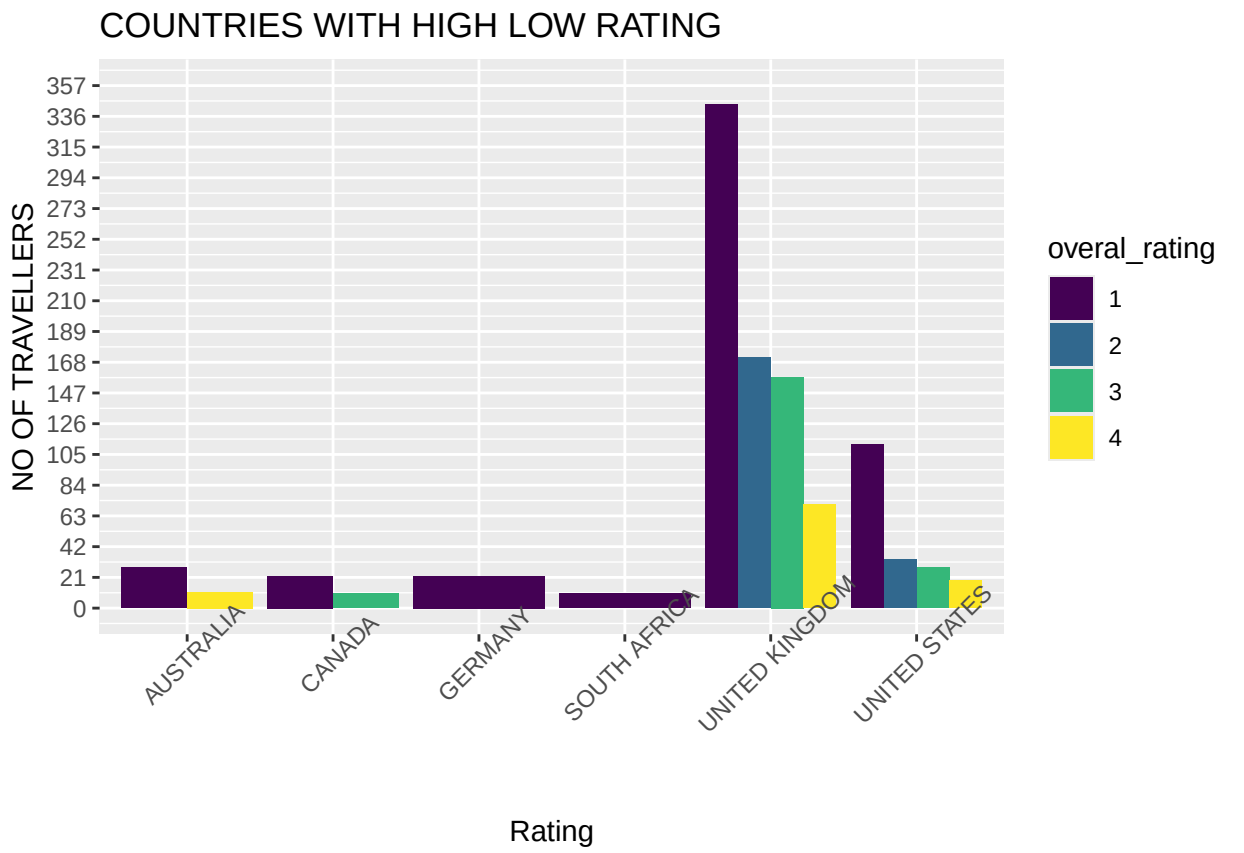
60% of customers of BRITISH AIRLINES are from UNITED KINGDOM followed distantly with 12.7% from UNITED STATES.

Low rating from 1 to 4

```
low_rating <- overall_verified %>%
    filter(overal_rating < 5)
```

```
country_lowrating_count <- low_rating %>%
    group_by(overal_rating) %>%
    count(review_country)

country_lowrating_count %>%
    filter(n > 9) %>%
    ggplot(aes(x = review_country,
        y = n, fill = overal_rating)) +
    geom_bar(stat = "identity",
        position = "dodge") +
    scale_y_continuous("NO OF TRAVELLERS",
        breaks = seq(0, 357, by = 21),
        limits = c(0, 357)) +
    theme(axis.text.x = element_text(angle = 45)) +
    labs(title = "COUNTRIES WITH HIGH LOW RATING",
        x = "Rating")
```



Many low rating scores were received from travellers from AUSTRALIA, CANADA, GERMAN, SOUTH AFRICA, UNITED KINGDOM and UNITED STATES.

UNITED KINGDOM had many travellers who had scored a 1.

Since the dataset has overal rating, it is unnecessary to perform sentiment analysis using review body and review title. It is also seen that the customers were rating individual category of service and provisions.

Thus, for the low rating we would connect with category reviews to get the different ratings that resulted in the overall low rating.

table_id in category_reviews is the id for the reviewer from the first reviewer to the 3000th reviewer, same as id column in low_rating data set.

1.4 Filtering low rating travellers who are in the category_reviews;

```
low_category <- category_reviews[category_reviews$table_id %in%
    low_rating$id, ]
low_category$table_id <- factor(low_category$table_id)
```

```r
head(low_category, 20)
```

```
# A tibble: 20 x 3
   category               score                          table_id
   <chr>                  <chr>                          <fct>
 1 Aircraft               A380                           4
 2 Type Of Traveller      Solo Leisure                   4
 3 Seat Type              Economy Class                  4
 4 Route                  Johannesburg to London Heathrow 4
 5 Date Flown             April 2025                     4
 6 Seat Comfort           1                              4
 7 Cabin Staff Service    1                              4
 8 Ground Service         1                              4
 9 Value For Money        1                              4
10 Recommended            no                             4
11 Type Of Traveller      Family Leisure                 8
12 Seat Type              Economy Class                  8
13 Route                  Dubai to London Heathrow       8
14 Date Flown             April 2025                     8
15 Seat Comfort           1                              8
16 Cabin Staff Service    1                              8
17 Food & Beverages       1                              8
18 Inflight Entertainment 1                              8
19 Ground Service         1                              8
20 Wifi & Connectivity    1                              8
```

```r
n_distinct(low_category$table_id)  ## no of distinct table_id
```

```
[1] 1271
```

We have 1271 distinct table_id implying that the data filtered was for the low rating customers.

Convert the data to a wide fortmat'

```r
low_category_2 <- low_category %>%
    pivot_wider(names_from = category, values_from = score)
```

```r
dim(low_category_2)
```

```
[1] 1271   14
```

```r
head(low_category_2, 20)
```

```
# A tibble: 20 x 14
   table_id Aircraft          `Type Of Traveller` `Seat Type` Route `Date Flown`
   <fct>    <chr>             <chr>               <chr>       <chr> <chr>
 1 4        A380              Solo Leisure        Economy Cl~ Joha~ April 2025
 2 8        <NA>              Family Leisure      Economy Cl~ Duba~ April 2025
 3 12       <NA>              Couple Leisure      Economy Cl~ Veni~ April 2025
 4 13       Airbus A321neo    Solo Leisure        Economy Cl~ Heat~ April 2025
 5 18       <NA>              Business            Business C~ Lond~ December 20~
 6 19       <NA>              Business            Premium Ec~ Lond~ January 2025
 7 20       <NA>              Business            Business C~ Cham~ January 2025
 8 24       A320              Solo Leisure        Economy Cl~ Lond~ February 20~
 9 25       <NA>              Business            Premium Ec~ Amst~ November 20~
10 26       A350-1000         Couple Leisure      Business C~ Lond~ February 20~
11 28       <NA>              Couple Leisure      Economy Cl~ Züri~ December 20~
12 32       <NA>              Solo Leisure        Economy Cl~ Manc~ November 20~
13 33       <NA>              Family Leisure      Premium Ec~ Hous~ December 20~
14 34       A320              Business            Economy Cl~ Lond~ January 2025
15 35       <NA>              Family Leisure      Economy Cl~ Lond~ November 20~
16 36       <NA>              Family Leisure      Economy Cl~ Lond~ January 2025
17 38       <NA>              Family Leisure      Economy Cl~ Larn~ November 20~
18 39       <NA>              Couple Leisure      Economy Cl~ Lond~ December 20~
19 40       Boeing 777 / A350 Business            Business C~ Wash~ December 20~
20 43       <NA>              Solo Leisure        Economy Cl~ Lond~ December 20~
# i 8 more variables: `Seat Comfort` <chr>, `Cabin Staff Service` <chr>,
#   `Ground Service` <chr>, `Value For Money` <chr>, Recommended <chr>,
#   `Food & Beverages` <chr>, `Inflight Entertainment` <chr>,
#   `Wifi & Connectivity` <chr>
```

Clean column names

```r
low_category_2 <- low_category_2 %>%
    clean_names()
```

Columns with missing values

```r
names(which(colSums(is.na(low_category_2)) > 0))
```

```
[1] "aircraft"              "type_of_traveller"     "route"
[4] "seat_comfort"          "cabin_staff_service"   "ground_service"
[7] "food_beverages"        "inflight_entertainment" "wifi_connectivity"
```

Number of missing values in each column

```
colSums(is.na(low_category_2))
```

```
             table_id              aircraft     type_of_traveller
                    0                   635                     2
            seat_type                 route           date_flown
                    0                     1                     0
         seat_comfort   cabin_staff_service        ground_service
                   80                    95                    46
      value_for_money           recommended        food_beverages
                    0                     0                   280
inflight_entertainment     wifi_connectivity
                  544                   951
```

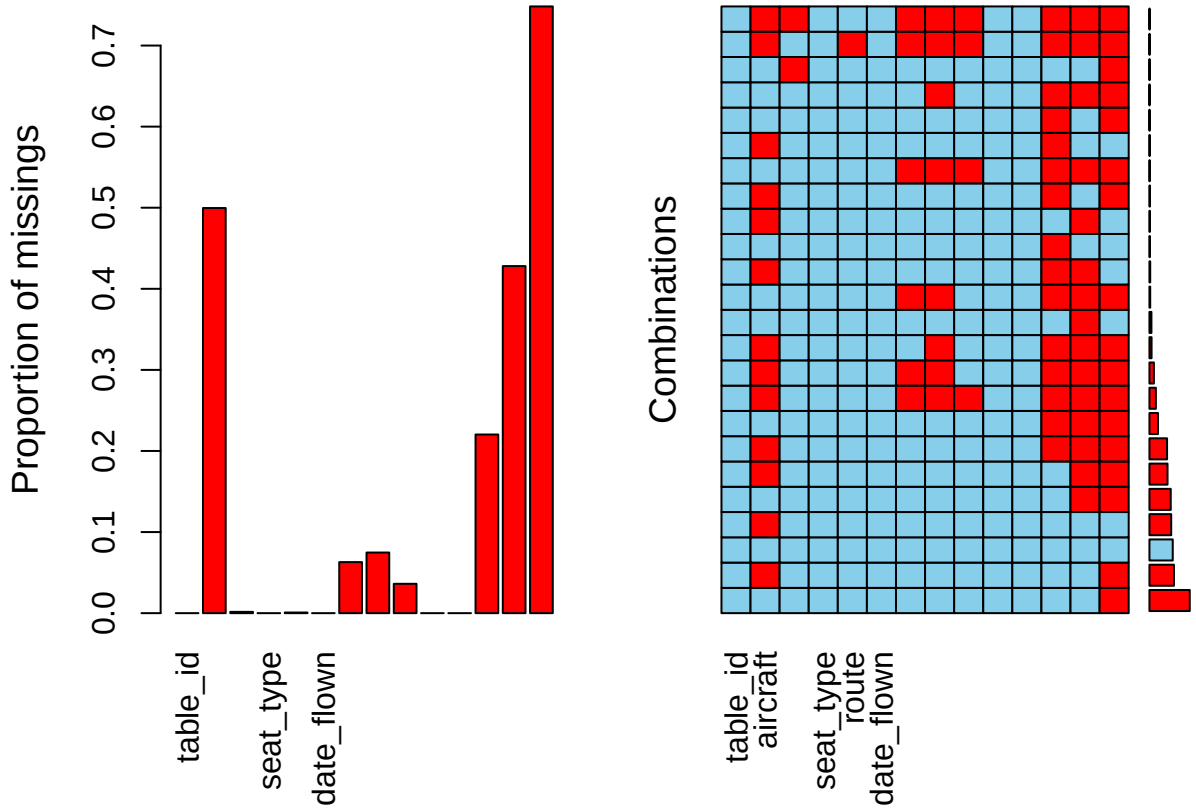
Columns without missing values

```
names(which(colSums(is.na(low_category_2)) == 0))
```

```
[1] "table_id"       "seat_type"      "date_flown"     "value_for_money"
[5] "recommended"
```

missing values

```
a <- aggr(low_category_2, plot = FALSE)
plot(a, numbers = TRUE)
```



It is seen that wifi connectivity has a very high proportion of missing values almost 75% of the entries are missing. Almost 50% of values on aircraft are missing while inflight_entertainment had more than 40%.

Given the high proportions of missing values in wifi_connectivity, inflight_entertainment, food_beverages and aircraft it will be hard to get meaningful insights from the columns as conclussions will be hard to draw. Thus, the columns will be removed;

```
low_category_3 <- low_category_2 %>%
    select(-wifi_connectivity, -inflight_entertainment,
        -aircraft, -food_beverages)
```

```
colMeans(is.na(low_category_3))
```

```
          table_id   type_of_traveller           seat_type              route
 0.000000000000000   0.001573564122738   0.000000000000000   0.000786782061369
        date_flown        seat_comfort cabin_staff_service     ground_service
 0.000000000000000   0.062942564909520   0.074744295830055   0.036191974822974
      value_for_money         recommended
 0.000000000000000   0.000000000000000
```

```r
head(low_category_3, 20)
```

```
# A tibble: 20 x 10
   table_id type_of_traveller seat_type       route      date_flown seat_comfort
   <fct>    <chr>             <chr>           <chr>      <chr>      <chr>
 1 4        Solo Leisure      Economy Class   Johannesb~ April 2025 1
 2 8        Family Leisure    Economy Class   Dubai to ~ April 2025 1
 3 12       Couple Leisure    Economy Class   Venice to~ April 2025 4
 4 13       Solo Leisure      Economy Class   Heathrow ~ April 2025 1
 5 18       Business          Business Class  London to~ December ~ 1
 6 19       Business          Premium Economy London to~ January 2~ 5
 7 20       Business          Business Class  Chambery ~ January 2~ 5
 8 24       Solo Leisure      Economy Class   London Ga~ February ~ 2
 9 25       Business          Premium Economy Amsterdam~ November ~ 3
10 26       Couple Leisure    Business Class  London to~ February ~ 5
11 28       Couple Leisure    Economy Class   Zürich to~ December ~ 2
12 32       Solo Leisure      Economy Class   Mancheste~ November ~ <NA>
13 33       Family Leisure    Premium Economy Houston t~ December ~ 1
14 34       Business          Economy Class   London to~ January 2~ 2
15 35       Family Leisure    Economy Class   London to~ November ~ 2
16 36       Family Leisure    Economy Class   London to~ January 2~ 1
17 38       Family Leisure    Economy Class   Larnaca t~ November ~ 1
18 39       Couple Leisure    Economy Class   London to~ December ~ 3
19 40       Business          Business Class  Washingto~ December ~ 3
20 43       Solo Leisure      Economy Class   London to~ December ~ 1
# i 4 more variables: cabin_staff_service <chr>, ground_service <chr>,
#   value_for_money <chr>, recommended <chr>
```

Type of travellers;

```r
low_category_3 %>%
    count(type_of_traveller)
```

```
# A tibble: 5 x 2
  type_of_traveller     n
  <chr>             <int>
1 Business            360
2 Couple Leisure      378
3 Family Leisure      169
4 Solo Leisure        362
5 <NA>                  2
```

Replace the NAs with Unknown for type of traveller;

```
low_category_3$type_of_traveller[is.na(low_category_3$type_of_traveller)] <- "Unknown"
```

```
low_category_3 %>%
    count(type_of_traveller, sort = T) %>%
    mutate(percent = n/sum(n) * 100)
```

```
# A tibble: 5 x 3
  type_of_traveller     n percent
  <chr>             <int>   <dbl>
1 Couple Leisure      378  29.7
2 Solo Leisure        362  28.5
3 Business            360  28.3
4 Family Leisure      169  13.3
5 Unknown               2   0.157
```

Most of the low rating customers were Couple Leisure, Solo Leisure and Business travellers. There numbers were also close.

Seat type;

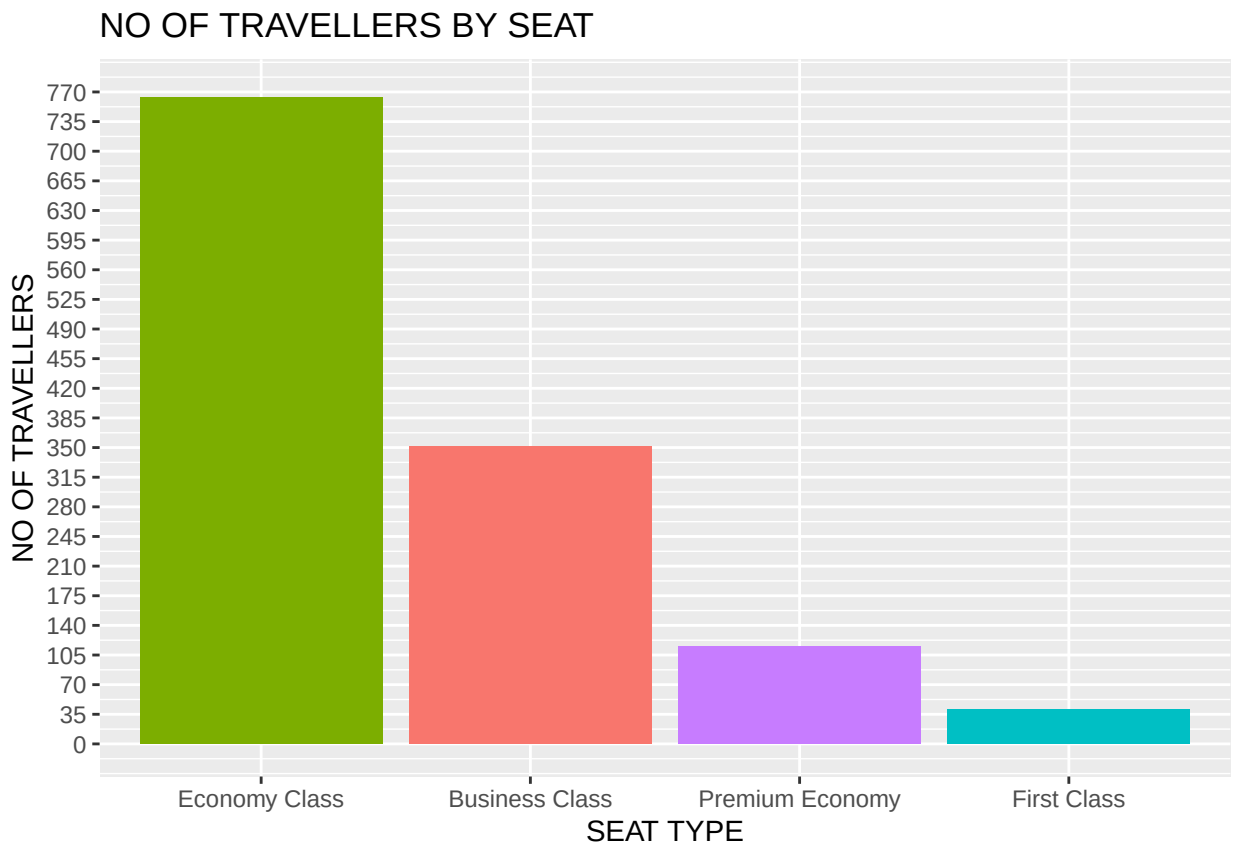
```
low_category_3 %>%
    count(seat_type, sort = T) %>%
    mutate(percent = n/sum(n) * 100)
```

```
# A tibble: 4 x 3
  seat_type           n percent
  <chr>           <int>   <dbl>
1 Economy Class     764   60.1
2 Business Class    351   27.6
3 Premium Economy   115    9.05
4 First Class        41    3.23
```

By far most of the low rating travellers used Economy Class;

```
low_category_3 %>%
    count(seat_type, sort = T) %>%
    ggplot(aes(reorder(x = seat_type, -n), y = n,
        fill = seat_type)) + geom_bar(stat = "identity",
    position = "dodge") + theme(legend.position = "none") +
    scale_y_continuous("NO OF TRAVELLERS", breaks = seq(0,
```

```
        770, by = 35), limits = c(0, 770)) +
    labs(title = "NO OF TRAVELLERS BY SEAT",
        x = "SEAT TYPE")
```



It is seen that more than 87% of the customers that gave a low rating used Economy Class and Business Class.

Seat comfort is a rating that ranges from 1 to 5;

```
range(low_category_3$seat_comfort)
```

```
[1] NA NA
```

```
low_category_3 %>%
    count(seat_comfort, sort = T) %>%
    mutate(percent = n/sum(n) * 100)
```

```
# A tibble: 6 x 3
  seat_comfort     n percent
```

```
  <chr>        <int>   <dbl>
1 2              371   29.2
2 1              346   27.2
3 5              268   21.1
4 3              180   14.2
5 <NA>            80    6.29
6 4               26    2.05
```

It is seen that seat_comfort ranges from 1 to 5 but it has NAs and it is loaded as a character column, we convert column to factor.

Since seat_comfort should be a fcator from 1 to 5 we will replace NAs with Unknown

```
low_category_4 <- low_category_3
low_category_4$seat_comfort <- factor(low_category_4$seat_comfort,
    levels = c(1, 2, 3, 4, 5), ordered = T)
low_category_4 <- low_category_4 %>%
    mutate(seat_comfort = fct_na_value_to_level(seat_comfort,
        level = "Unknown"))
```

```
range(low_category_4$seat_comfort)
```

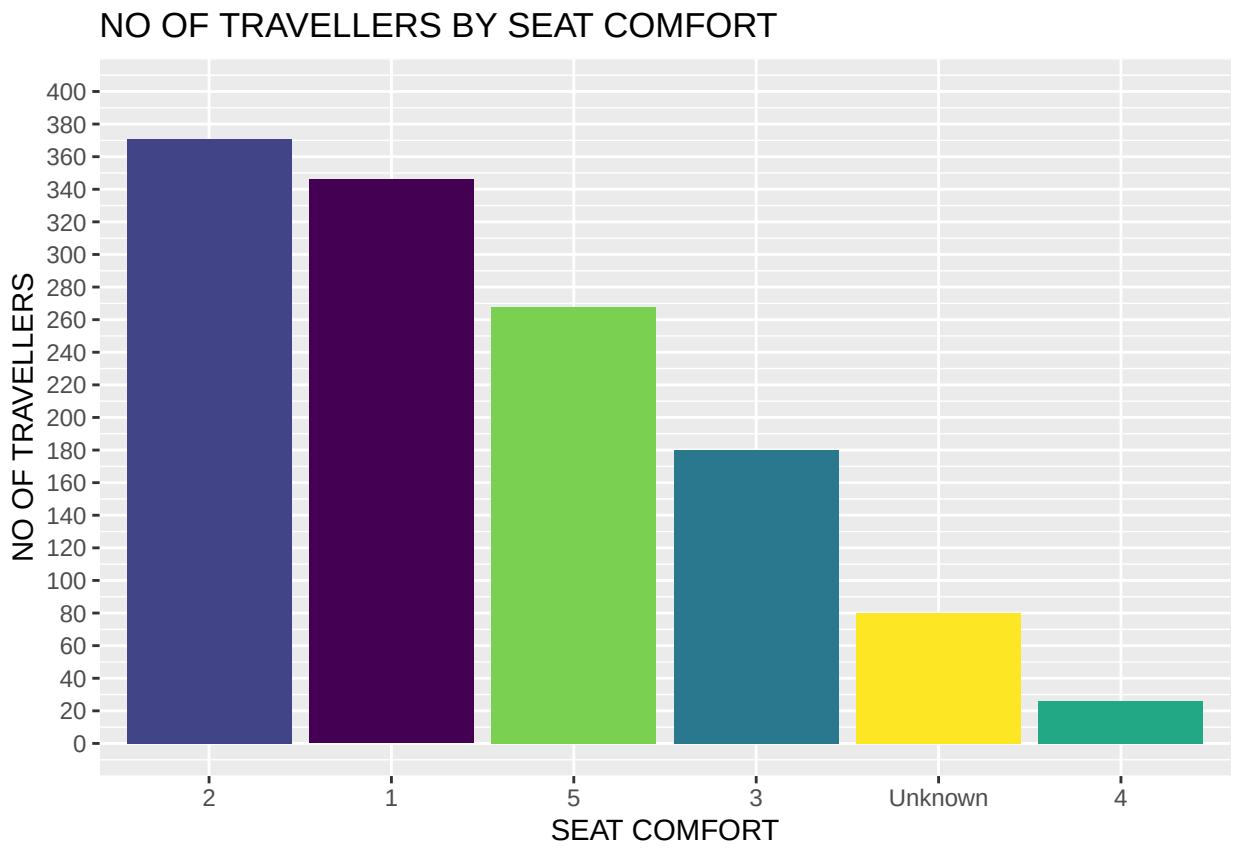
```
[1] 1       Unknown
Levels: 1 < 2 < 3 < 4 < 5 < Unknown
```

```
low_category_4 %>%
    count(seat_comfort, sort = T) %>%
    mutate(percent = n/sum(n) * 100)
```

```
# A tibble: 6 x 3
  seat_comfort     n percent
  <ord>        <int>   <dbl>
1 2              371   29.2
2 1              346   27.2
3 5              268   21.1
4 3              180   14.2
5 Unknown         80    6.29
6 4               26    2.05
```

```
low_category_4 %>%
    count(seat_comfort, sort = T) %>%
    ggplot(aes(reorder(x = seat_comfort,
        -n), y = n, fill = seat_comfort)) +
    geom_bar(stat = "identity",
```

```
        position = "dodge") +
    theme(legend.position = "none") +
    scale_y_continuous("NO OF TRAVELLERS",
        breaks = seq(0, 400, by = 20),
        limits = c(0, 400)) +
    labs(title = "NO OF TRAVELLERS BY SEAT COMFORT",
        x = "SEAT COMFORT")
```



21% of travellers that gave a low rating did not have a problem with their seat although at least 75% of them had a problem with their seat as they rated seat comfort with 2,1 and 3.

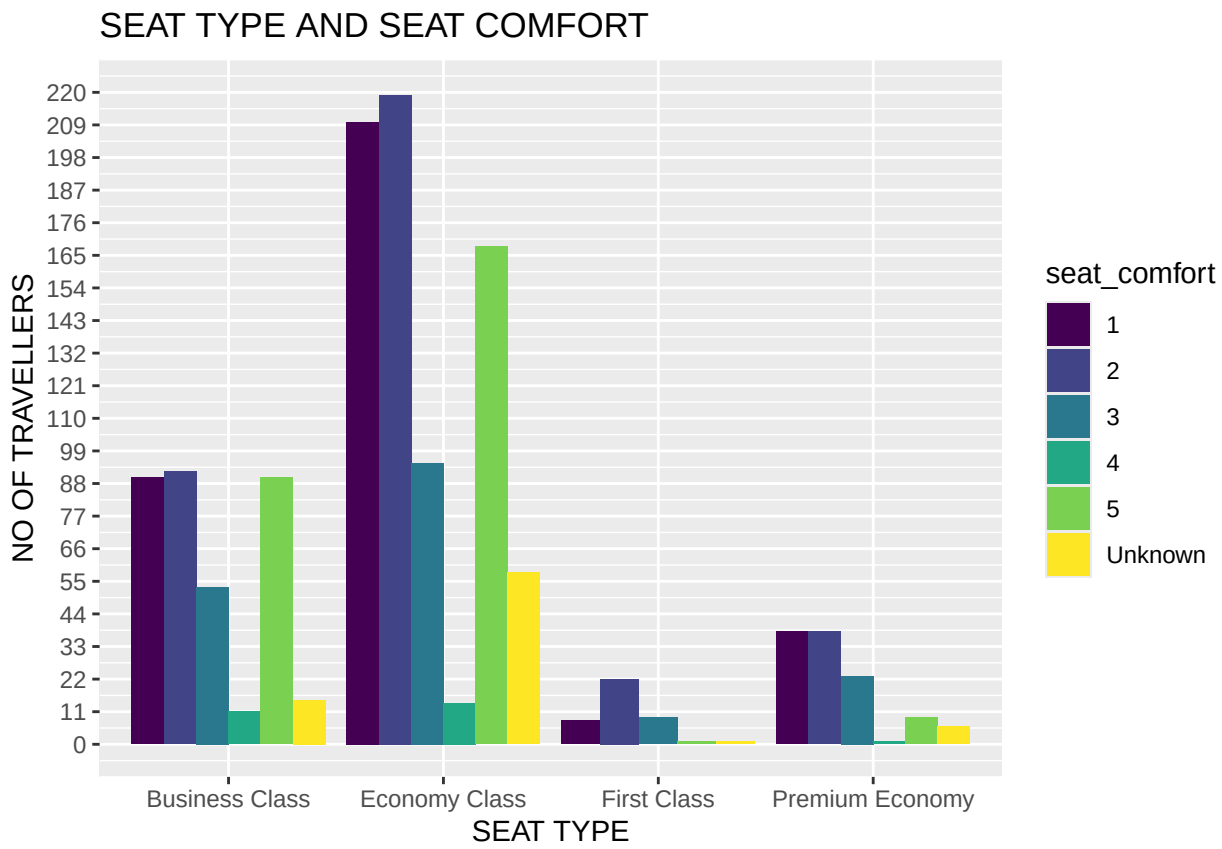
```
seat_type_comfort <- low_category_4 %>%
    group_by(seat_type) %>%
    count(seat_comfort)
seat_type_comfort_2 <- low_category_4 %>%
    group_by(seat_type) %>%
    count(seat_comfort) %>%
    mutate(percent = n/sum(n) * 100)
```

```
knitr::kable(seat_type_comfort_2, caption = "SEAT TYPE AND SEAT COMFORT",
    longtable = TRUE, digits = 2, format.args = list(big.mark = ",",
        scientific = FALSE), "latex") %>%
    kableExtra::column_spec(1, border_left = T) %>%
    kableExtra::column_spec(4, border_right = T) %>%
    kableExtra::kable_styling(latex_options = "HOLD_position",
        "repeat_header")
```

Table 5: SEAT TYPE AND SEAT COMFORT

seat_type	seat_comfort	n	percent
Business Class	1	90	25.64
Business Class	2	92	26.21
Business Class	3	53	15.10
Business Class	4	11	3.13
Business Class	5	90	25.64
Business Class	Unknown	15	4.27
Economy Class	1	210	27.49
Economy Class	2	219	28.66
Economy Class	3	95	12.43
Economy Class	4	14	1.83
Economy Class	5	168	21.99
Economy Class	Unknown	58	7.59
First Class	1	8	19.51
First Class	2	22	53.66
First Class	3	9	21.95
First Class	5	1	2.44
First Class	Unknown	1	2.44
Premium Economy	1	38	33.04
Premium Economy	2	38	33.04
Premium Economy	3	23	20.00
Premium Economy	4	1	0.87
Premium Economy	5	9	7.83
Premium Economy	Unknown	6	5.22

```
ggplot(seat_type_comfort, aes(x = seat_type,
    y = n, fill = seat_comfort)) +
    geom_bar(stat = "identity",
        position = "dodge") +
    scale_y_continuous("NO OF TRAVELLERS",
        breaks = seq(0, 220, by = 11),
        limits = c(0, 220)) +
    labs(title = "SEAT TYPE AND SEAT COMFORT",
        x = "SEAT TYPE")
```



Business and Economy had high numbers of travellers who were comfortable with their seats while Premium Economy and First Class were unhappy with their seats. Overally, most of the travellers were not comfortable with their seats across the 4 seat types.

Cabin staff service;

```
low_category_4 %>%
    count(cabin_staff_service, sort = T) %>%
    mutate(percent = n/sum(n) * 100)
```

```
# A tibble: 7 x 3
  cabin_staff_service     n percent
  <chr>               <int>   <dbl>
1 2                     441  34.7
2 1                     327  25.7
3 3                     266  20.9
4 <NA>                   95   7.47
5 4                      93   7.32
6 0                      37   2.91
7 5                      12   0.944
```

Cabin staff service ranges from 0 to 5, but it has 95 NAs.

We convert the column to an ordered factor and replace NAs with unknown.

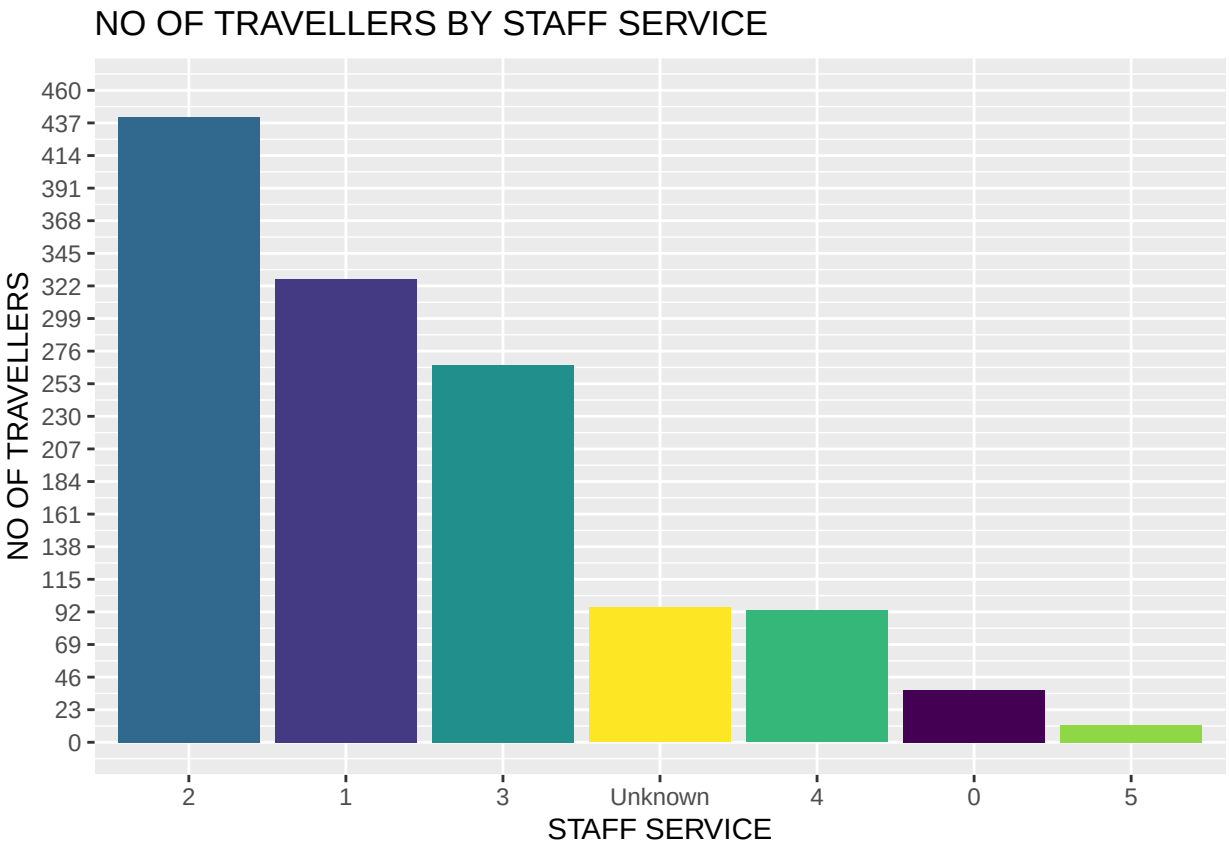
```
low_category_4$cabin_staff_service <- factor(low_category_4$cabin_staff_service,
    levels = c(0, 1, 2, 3, 4, 5), ordered = T)
low_category_4 <- low_category_4 %>%
    mutate(cabin_staff_service = fct_na_value_to_level(cabin_staff_service,
        level = "Unknown"))
```

```
low_category_4 %>%
    count(cabin_staff_service, sort = T) %>%
    mutate(percent = n/sum(n) * 100)
```

```
# A tibble: 7 x 3
  cabin_staff_service     n percent
  <ord>               <int>   <dbl>
1 2                     441  34.7
2 1                     327  25.7
3 3                     266  20.9
4 Unknown                95   7.47
5 4                      93   7.32
6 0                      37   2.91
7 5                      12   0.944
```

At least 81% of travellers that gave an overall low rating had also rated cabin staff service lowly.

```
low_category_4 %>%
    count(cabin_staff_service,
        sort = T) %>%
    ggplot(aes(reorder(x = cabin_staff_service,
        -n), y = n, fill = cabin_staff_service)) +
    geom_bar(stat = "identity",
        position = "dodge") +
    theme(legend.position = "none") +
    scale_y_continuous("NO OF TRAVELLERS",
        breaks = seq(0, 460, by = 23),
        limits = c(0, 460)) +
    labs(title = "NO OF TRAVELLERS BY STAFF SERVICE",
        x = "STAFF SERVICE")
```



Staff Service per Seat type

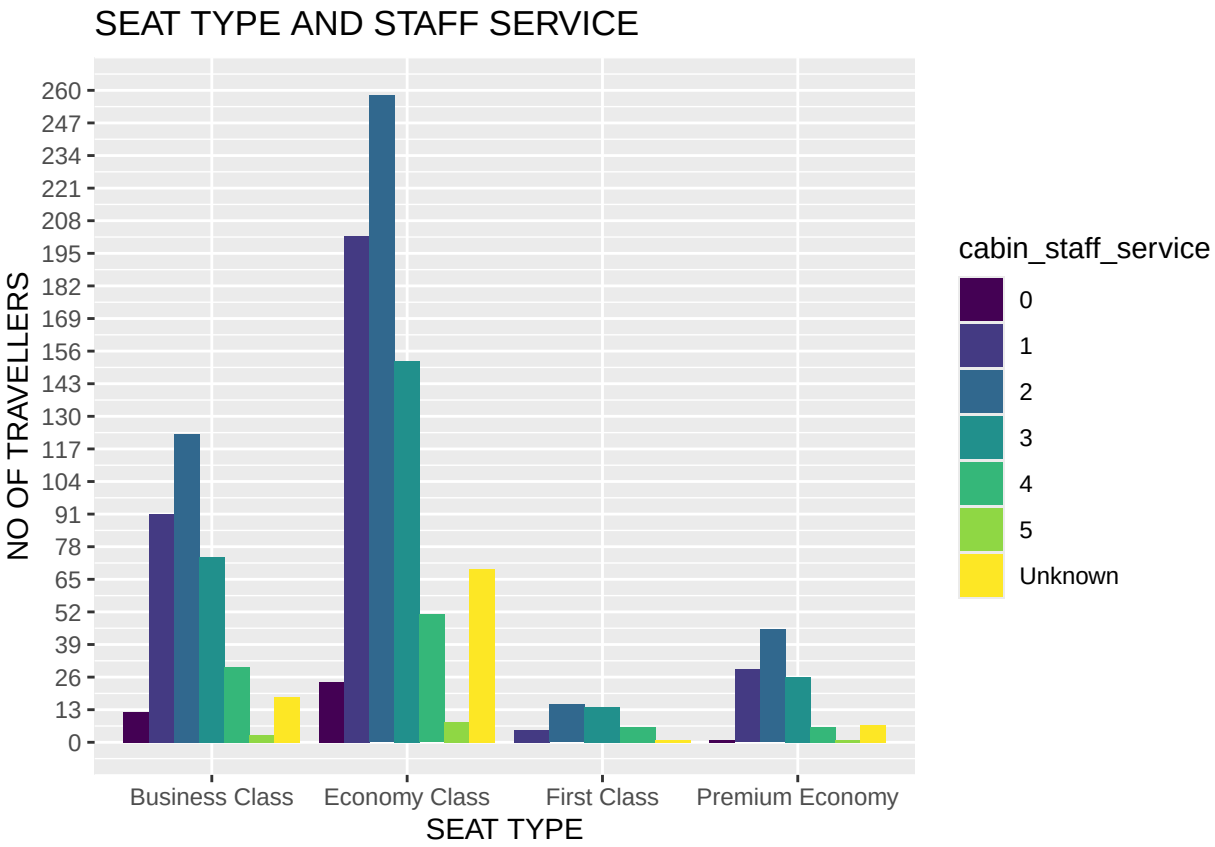
```
seat_type_service <- low_category_4 %>%
    group_by(seat_type) %>%
    count(cabin_staff_service)
seat_type_service_2 <- low_category_4 %>%
    group_by(seat_type) %>%
    count(cabin_staff_service) %>%
    mutate(percent = n/sum(n) * 100)
```

```
knitr::kable(seat_type_service_2, caption = "SEAT TYPE AND STAFF SERVICE",
    longtable = TRUE, digits = 2, format.args = list(big.mark = ",",
        scientific = FALSE), "latex") %>%
    kableExtra::column_spec(1, border_left = T) %>%
    kableExtra::column_spec(4, border_right = T) %>%
    kableExtra::kable_styling(latex_options = "HOLD_position",
        "repeat_header")
```

Table 6: SEAT TYPE AND STAFF SERVICE

seat_type	cabin_staff_service	n	percent
Business Class	0	12	3.42
Business Class	1	91	25.93
Business Class	2	123	35.04
Business Class	3	74	21.08
Business Class	4	30	8.55
Business Class	5	3	0.85
Business Class	Unknown	18	5.13
Economy Class	0	24	3.14
Economy Class	1	202	26.44
Economy Class	2	258	33.77
Economy Class	3	152	19.90
Economy Class	4	51	6.68
Economy Class	5	8	1.05
Economy Class	Unknown	69	9.03
First Class	1	5	12.20
First Class	2	15	36.59
First Class	3	14	34.15
First Class	4	6	14.63
First Class	Unknown	1	2.44
Premium Economy	0	1	0.87
Premium Economy	1	29	25.22
Premium Economy	2	45	39.13
Premium Economy	3	26	22.61
Premium Economy	4	6	5.22
Premium Economy	5	1	0.87
Premium Economy	Unknown	7	6.09

```
ggplot(seat_type_service, aes(x = seat_type,
    y = n, fill = cabin_staff_service)) +
    geom_bar(stat = "identity",
        position = "dodge") +
    scale_y_continuous("NO OF TRAVELLERS",
        breaks = seq(0, 260, by = 13),
        limits = c(0, 260)) +
    labs(title = "SEAT TYPE AND STAFF SERVICE",
        x = "SEAT TYPE")
```



for the seat type and cabin staff service we make a rating higher than 3 to be high that is 4 and 5 and the rest to be low.

```
seat_type_service_2$score <- ifelse(seat_type_service_2$cabin_staff_service >
    3, "high", "low")
```

We maintain the value of unknown to be unknown

```
seat_type_service_2 <- seat_type_service_2 %>%
    mutate(score = case_when(str_detect(cabin_staff_service,
        "Unknown") ~ "Unknown", TRUE ~ score))
```

total percentage per class of seat type and cabin staff service

```
seat_type_service_3 <- seat_type_service_2 %>%
    group_by(seat_type, score) %>%
    summarise(percent = sum(percent))
```

```
knitr::kable(seat_type_service_3, caption = "SEAT TYPE AND STAFF SERVICE TOTALS",
    longtable = TRUE, digits = 2, format.args = list(big.mark = ",",
        scientific = FALSE), "latex") %>%
    kableExtra::column_spec(1, border_left = T) %>%
    kableExtra::column_spec(3, border_right = T) %>%
    kableExtra::kable_styling(latex_options = "HOLD_position",
        "repeat_header")
```

Table 7: SEAT TYPE AND STAFF SERVICE TOTALS

seat_type	score	percent
Business Class	Unknown	5.13
Business Class	high	9.40
Business Class	low	85.47
Economy Class	Unknown	9.03
Economy Class	high	7.72
Economy Class	low	83.25
First Class	Unknown	2.44
First Class	high	14.63
First Class	low	82.93
Premium Economy	Unknown	6.09
Premium Economy	high	6.09
Premium Economy	low	87.83

It seen that low rating for cabin staff service was maintained across the different seat types.

Ground Service;

```
low_category_4 %>%
    count(ground_service, sort = T) %>%
    mutate(percent = n/sum(n) * 100)
```

```
# A tibble: 7 x 3
  ground_service     n percent
  <chr>          <int>   <dbl>
1 1                452 35.6
2 2                393 30.9
3 3                191 15.0
4 0                174 13.7
5 <NA>              46  3.62
6 4                 14  1.10
7 5                  1  0.0787
```

Ground Service is rated from 0 to 5 but has 46 NAs values.

We convert the column to factor and replace NAs with unknown;

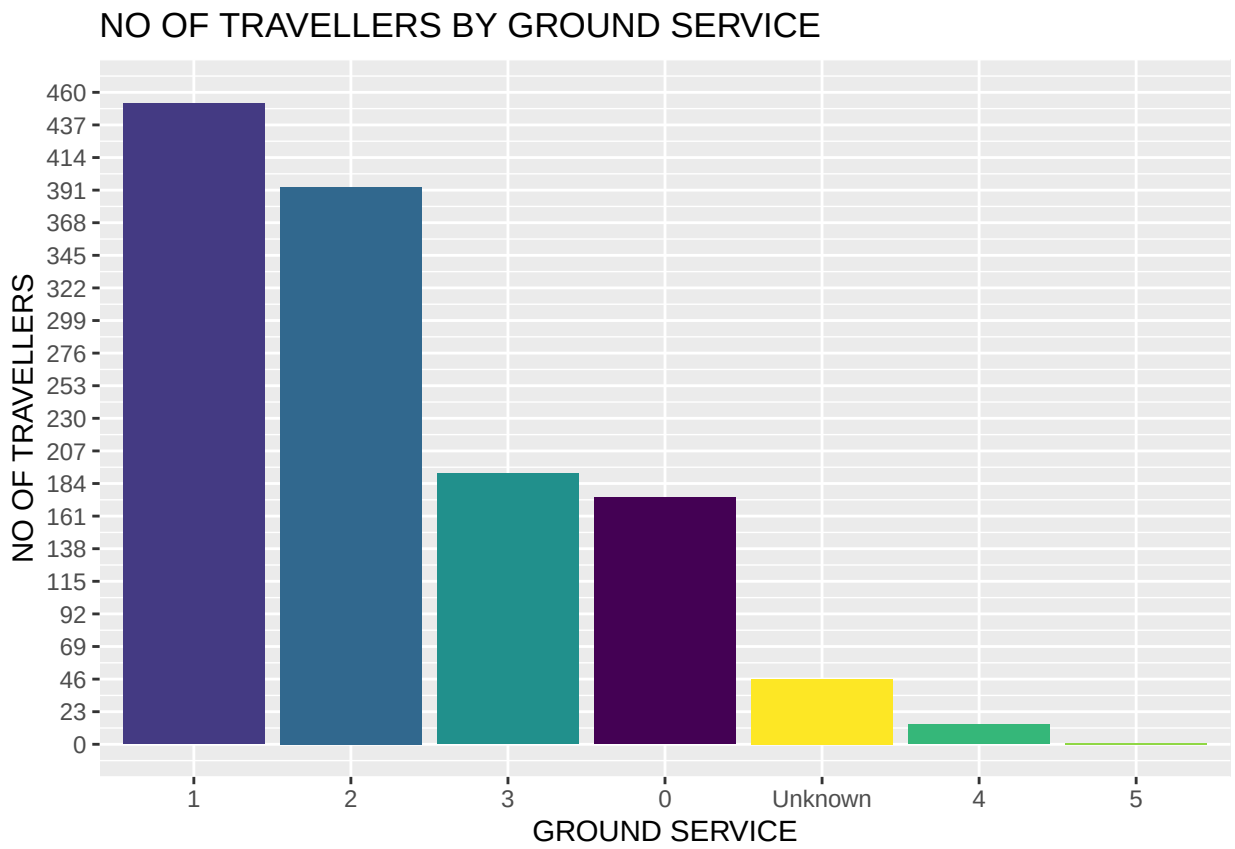
```
low_category_4$ground_service <- factor(low_category_4$ground_service,
    levels = c(0, 1, 2, 3, 4, 5), ordered = T)
low_category_4 <- low_category_4 %>%
    mutate(ground_service = fct_na_value_to_level(ground_service,
        level = "Unknown"))
```

```
low_category_4 %>%
    count(ground_service, sort = T) %>%
    mutate(percent = n/sum(n) * 100)
```

```
# A tibble: 7 x 3
  ground_service     n percent
  <ord>          <int>   <dbl>
1 1                452 35.6
2 2                393 30.9
3 3                191 15.0
4 0                174 13.7
5 Unknown           46  3.62
6 4                 14  1.10
7 5                  1  0.0787
```

At least 95% of the travellers that gave an overall low rating were not happy with the ground service.

```
low_category_4 %>%
    count(ground_service, sort = T) %>%
    ggplot(aes(reorder(x = ground_service,
        -n), y = n, fill = ground_service)) +
    geom_bar(stat = "identity",
        position = "dodge") +
    theme(legend.position = "none") +
    scale_y_continuous("NO OF TRAVELLERS",
        breaks = seq(0, 460, by = 23),
        limits = c(0, 460)) +
    labs(title = "NO OF TRAVELLERS BY GROUND SERVICE",
        x = "GROUND SERVICE")
```



Ground Service per Seat type

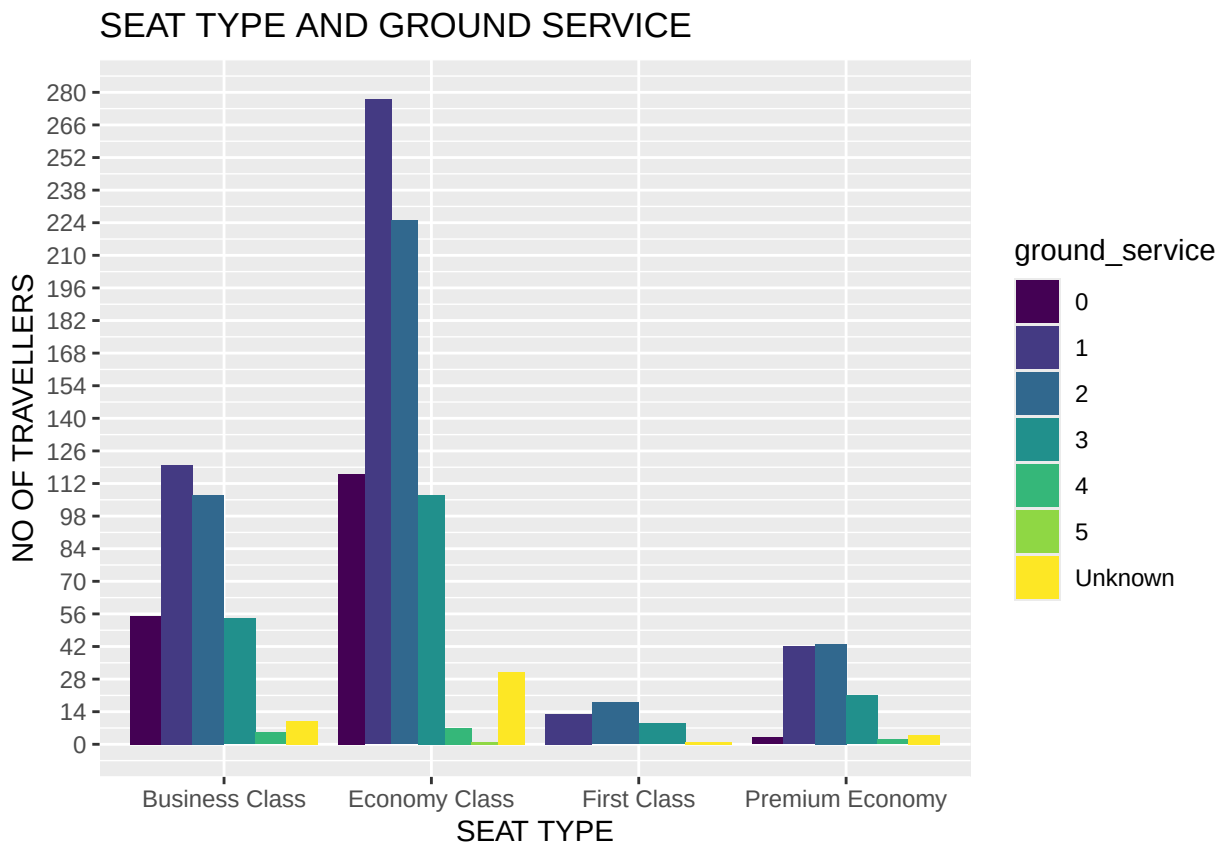
```
seat_type_ground <- low_category_4 %>%
    group_by(seat_type) %>%
    count(ground_service)
seat_type_ground_2 <- low_category_4 %>%
    group_by(seat_type) %>%
    count(ground_service) %>%
    mutate(percent = n/sum(n) * 100)
```

```
knitr::kable(seat_type_ground_2, caption = "SEAT TYPE AND GROUND SERVICE",
    longtable = TRUE, digits = 2, format.args = list(big.mark = ",",
        scientific = FALSE), "latex") %>%
    kableExtra::column_spec(1, border_left = T) %>%
    kableExtra::column_spec(4, border_right = T) %>%
    kableExtra::kable_styling(latex_options = "HOLD_position",
        "repeat_header")
```

Table 8: SEAT TYPE AND GROUND SERVICE

seat_type	ground_service	n	percent
Business Class	0	55	15.67
Business Class	1	120	34.19
Business Class	2	107	30.48
Business Class	3	54	15.38
Business Class	4	5	1.42
Business Class	Unknown	10	2.85
Economy Class	0	116	15.18
Economy Class	1	277	36.26
Economy Class	2	225	29.45
Economy Class	3	107	14.01
Economy Class	4	7	0.92
Economy Class	5	1	0.13
Economy Class	Unknown	31	4.06
First Class	1	13	31.71
First Class	2	18	43.90
First Class	3	9	21.95
First Class	Unknown	1	2.44
Premium Economy	0	3	2.61
Premium Economy	1	42	36.52
Premium Economy	2	43	37.39
Premium Economy	3	21	18.26
Premium Economy	4	2	1.74
Premium Economy	Unknown	4	3.48

```
ggplot(seat_type_ground_2, aes(x = seat_type,
    y = n, fill = ground_service)) +
    geom_bar(stat = "identity",
        position = "dodge") +
    scale_y_continuous("NO OF TRAVELLERS",
        breaks = seq(0, 280, by = 14),
        limits = c(0, 280)) +
    labs(title = "SEAT TYPE AND GROUND SERVICE",
        x = "SEAT TYPE")
```



The rating was low across the different seat types.

for the seat type and ground service we make a rating higher than 3 to be high that is 4 and 5 and the rest to be low.

```
seat_type_ground_2$score <- ifelse(seat_type_ground_2$ground_service >
    3, "high", "low")
```

We maintain the value of unknown to be unknown

```
seat_type_ground_2 <- seat_type_ground_2 %>%
    mutate(score = case_when(str_detect(ground_service,
        "Unknown") ~ "Unknown", TRUE ~ score))
```

total percentage per class of seat type and ground service

```r
seat_type_ground_3 <- seat_type_ground_2 %>%
    group_by(seat_type, score) %>%
    summarise(percent = sum(percent))
```

```r
knitr::kable(seat_type_ground_3, caption = "SEAT TYPE AND GROYUND SERVICE TOTALS",
    longtable = TRUE, digits = 2, format.args = list(big.mark = ",",
        scientific = FALSE), "latex") %>%
    kableExtra::column_spec(1, border_left = T) %>%
    kableExtra::column_spec(3, border_right = T) %>%
    kableExtra::kable_styling(latex_options = "HOLD_position",
        "repeat_header")
```

Table 9: SEAT TYPE AND GROYUND SERVICE TOTALS

seat_type	score	percent
Business Class	Unknown	2.85
Business Class	high	1.42
Business Class	low	95.73
Economy Class	Unknown	4.06
Economy Class	high	1.05
Economy Class	low	94.90
First Class	Unknown	2.44
First Class	low	97.56
Premium Economy	Unknown	3.48
Premium Economy	high	1.74
Premium Economy	low	94.78

Low rating of ground service is high across the different seat types for the overal low rating travellers.

Value for Money;

```r
low_category_4 %>%
    count(value_for_money, sort = T) %>%
    mutate(percent = n/sum(n) * 100)
```

```
# A tibble: 5 x 3
  value_for_money     n percent
  <chr>           <int>   <dbl>
1 1                 448   35.2
2 2                 375   29.5
3 0                 247   19.4
4 3                 174   13.7
5 4                  27    2.12
```

Value for money is given a rating from 0 to 4. We convert the column to factor with ordered levels from 0 to 4;

```
low_category_4$value_for_money <- factor(low_category_4$value_for_money,
    levels = c(0, 1, 2, 3, 4), ordered = T)
```

```
low_category_4 %>%
    count(value_for_money, sort = T) %>%
    mutate(percent = n/sum(n) * 100)
```

```
# A tibble: 5 x 3
  value_for_money     n percent
  <ord>           <int>   <dbl>
1 1                 448   35.2
2 2                 375   29.5
3 0                 247   19.4
4 3                 174   13.7
5 4                  27    2.12
```

84% of travellers with a low rating had a rating of 0, 1 or 2 on value for money which can be rated as low.

```
low_category_4 %>%
    count(value_for_money, sort = T) %>%
    ggplot(aes(reorder(x = value_for_money,
        -n), y = n, fill = value_for_money)) +
    geom_bar(stat = "identity",
        position = "dodge") +
    theme(legend.position = "none") +
    scale_y_continuous("NO OF TRAVELLERS",
        breaks = seq(0, 460, by = 23),
        limits = c(0, 460)) +
    labs(title = "NO OF TRAVELLERS BY VALUE FOR MONEY",
        x = "VALUE FOR MONEY")
```

Would the travellers recommend the airline;

```
low_category_4 %>%
    count(recommended, sort = T) %>%
    mutate(percent = n/sum(n) * 100)
```

```
# A tibble: 2 x 3
  recommended     n percent
  <chr>       <int>   <dbl>
1 no           1256   98.8
2 yes            15    1.18
```

At least 98% of travellers with overal low rating would not recommend the Airline

- The data was scrapped in two different datasets

i. overal_reviews- data set that had 3350 observations with 8 variables. The data set had the name of reviewer, date of review, overal rating, trip occurrence-trip verification, review body-comment, review title, review country and id-which was the id given as scrapping was done from the first reviewer.

ii. category_reviews- data set that had 33629 observations with three variables of category-had observations like type of traveler, seat type, route, date flown, seat comfort, cabin staff service, ground service, Value For Money, Aircraft, Food & Beverages, Inflight Entertainment, Wifi & Connectivity and Recommended, score-was the value taken category entries and table_id-which was the id of the first reviewer to the last reviewer, it was recycled from 1 to 3000 implying the first 3000 reviewers.

Reviews were from 04/03/2015 to 18/05/2025.

- Out of the 3350 reviews in the overal reviews 2098 were verified travelers, that is 63% of the reviews scrapped were of verified travelers.
- There were 173-8%, individuals who were verified to have traveled more than once with the most frequent travelers having traveled 34 times, followed by 28 times, then 13 times and 2 individuals having traveled 10 times. It is worth noting that 102 of the 173 travelers only used British Airways twice.
- Overall rating ranged from 1 to 10. The column was converted to an ordered factor column with 1 as the lowest level and 10 as the highest level.
- Out of the 2098 verified travelers 1271 which is 61% of the travelers, gave a rating of 4 and below. At least 29% of them gave a rating of 1, the lowest rating.
- 60% of customers of BRITISH AIRLINES are from UNITED KINGDOM followed distantly with 12.7% from UNITED STATES.
- Analysis was further done on travelers that gave a low rating of 1 to 4.
- Many low rating scores were received from travelers from AUSTRALIA, CANADA, GERMANY, SOUTH AFRICA, UNITED KINGDOM and UNITED STATES.
- UNITED KINGDOM had many travelers who had scored a 1.
- A data set of low rating was created and it was joined with the category reviews by table_id and id. We got 1271 distinct table_id.
- Obtained data was converted to wide format where columns entries were converted to column heads.
- A data set of 1271 observations with 14 variables was obtained.
- The data set had 9 columns with missing values where aircraft, food_beverages, inflight_entertainment and wifi_connectivity had had high number of missing values. The columns were removed.
- NAs in type_of_traveller were replaced by unknown.
- Most of the low rating customers were Couple Leisure, Solo Leisure and Business travelers.There numbers were also close.
- By far most of the low rating travelers that is 60%, used Economy Class.
- At least 87% of the customers that gave a low rating used Economy Class and Business Class.

- seat comfort column was converted to factor with an ordered levels from 1 to 5 with NAs replaced with Unknown.
- 21% of travelers that gave a low rating did not have a problem with their seat although at least 75% of them had a problem with their seat as they rated seat comfort at 2, 1 and 3.
- Business and Economy had high numbers of travelers who were comfortable with their seats while Premium Economy and First Class were unhappy with their seats although, most of the travelers were not comfortable with their seats across the 4 seat types.
- Cabin staff service column was converted to factor with an ordered levels from 0 to 5 with NAs replaced with Unknown.
- At least 81% of travelers that gave an overall low rating had also rated cabin staff service lowly. low rating for cabin staff service was maintained across the different seat types.
- At least 95% of the travelers that gave an overall low rating were not happy with the ground service. Low rating of ground service is high across the different seat types for the overall low rating travelers.
- 84% of travelers with a low rating had a rating of 0, 1 or 2 on value for money which can be rated as low.
- At least 98% of travelers with overall low rating would not recommend the Airline.