Customer Analysis

Summary

This report includes sections with:

- 1. Background
- 2. Questions
- 3. Exploratory data analysis
- 4. References

1. Background

Customer churn or customer attrition: Loss of clients or when customers stop doing business with a service or company.

Tenure: A measure for how long the client/customer stays

2. Questions

What do we know about the data?

Information from the article **Using customer behavior data to improve customer retention**

Data Quality and Validity

1. Where does the data come from? How was data collected?
2. Was data collected with any prior objective/purpose?
3. Is data suitable for the *Question*? (The company's problem/issue/questions)
4. Are there missing values in the data?

Data Analysis questions

"A telecommunications company is concerned about revenue and the number of costumers leaving their landline business for cable competitors. They need to understand who is leaving. Imagine that we are analysts at this company and we have to find out who is leaving and why"

1. What is the percentage of churn?
2. What can possibly cause the churn? What are the variables associated with churn?
3. How does costumer behavior vary and how does it influence the whole analysis?

3. Exploratory data analysis

3.1. Libraries

```
library(tibble)
library(readr)
library(summarytools)
library(tidyverse)
```

3.2. Read data

Data source: Telecom dataset from IBM Sample Data Sets

Telco costumer data set

```
churn <- read.csv("data/Telco_Customer_Churn.csv")
```

```
# With read_csv from readr doens't convert variables into factors
churn_dat <- readr::read_csv("data/Telco_Customer_Churn.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   SeniorCitizen = col_integer(),
##   tenure = col_integer(),
##   MonthlyCharges = col_double(),
##   TotalCharges = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
tibble::glimpse(churn_dat)
```

```
## Observations: 7,043
## Variables: 21
## $ customerID       <chr> "7590-VHVEG", "5575-GNVDE", "3668-QPYBK", "77...
## $ gender           <chr> "Female", "Male", "Male", "Male", "Female", "...
## $ SeniorCitizen    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ Partner          <chr> "Yes", "No", "No", "No", "No", "No", "No", "N...
## $ Dependents       <chr> "No", "No", "No", "No", "No", "No", "Yes", "N...
## $ tenure           <int> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 5...
## $ PhoneService     <chr> "No", "Yes", "Yes", "No", "Yes", "Yes", "Yes"...
## $ MultipleLines    <chr> "No phone service", "No", "No", "No phone ser...
## $ InternetService  <chr> "DSL", "DSL", "DSL", "DSL", "Fiber optic", "F...
## $ OnlineSecurity   <chr> "No", "Yes", "Yes", "Yes", "No", "No", "No", ...
## $ OnlineBackup     <chr> "Yes", "No", "Yes", "No", "No", "No", "Yes", ...
## $ DeviceProtection <chr> "No", "Yes", "No", "Yes", "No", "Yes", "No", ...
## $ TechSupport      <chr> "No", "No", "No", "Yes", "No", "No", "No", "N...
## $ StreamingTV      <chr> "No", "No", "No", "No", "No", "Yes", "Yes", "...
## $ StreamingMovies  <chr> "No", "No", "No", "No", "No", "Yes", "No", "N...
## $ Contract         <chr> "Month-to-month", "One year", "Month-to-month...
## $ PaperlessBilling <chr> "Yes", "No", "Yes", "No", "Yes", "Yes", "Yes"...
## $ PaymentMethod    <chr> "Electronic check", "Mailed check", "Mailed c...
## $ MonthlyCharges   <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89....
## $ TotalCharges     <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820....
## $ Churn            <chr> "No", "No", "Yes", "No", "Yes", "Yes", "No", ...
```

3.3. Understand data

Data variables meaning

- Churn: Customers who left within the last month
- *Services that each customer has signed up for - phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies*
- *Customer account information - how long they've been a customer, contract type, payment method, paperless billing, monthly charges, and total charges*
- *Demographic info about customers - gender, age range, and if they have partners and dependents*

```
summarytools::descr(churn_dat)
```

```
## Non-numerical variable(s) ignored: customerID, gender, Partner, Dependents, PhoneService, MultipleLi
```

```
## Descriptive Statistics
## Data Frame: churn_dat
## N: 7043
##
##                     SeniorCitizen    tenure   MonthlyCharges   TotalCharges
## ----------------- --------------- --------- ---------------- --------------
##              Mean            0.16     32.37            64.76        2283.30
##           Std.Dev            0.37     24.56            30.09        2266.77
##               Min            0.00      0.00            18.25          18.80
##            Median            0.00     29.00            70.35        1397.47
##               Max            1.00     72.00           118.75        8684.80
##               MAD            0.00     32.62            35.66        1812.92
##               IQR            0.00     46.00            54.35        3393.29
##                CV            0.44      1.32             2.15           1.01
##          Skewness            1.83      0.24            -0.22           0.96
##       SE.Skewness            0.03      0.03             0.03           0.03
##          Kurtosis            1.36     -1.39            -1.26          -0.23
##           N.Valid         7043.00   7043.00          7043.00        7032.00
##         Pct.Valid          100.00    100.00           100.00          99.84
```

3.4 Data cleaning

We introduce some data cleaning here after realising variable SeniorCitizen is a dichotomic variable that should be converted to character type, with "yes" and "no" values.

```
head(churn_dat)
```

```
## # A tibble: 6 x 21
##   customerID gender SeniorCitizen Partner Dependents tenure PhoneService
##        <chr>  <chr>         <int>   <chr>      <chr>  <int>        <chr>
## 1 7590-VHVEG Female             0     Yes         No      1           No
## 2 5575-GNVDE   Male             0      No         No     34          Yes
## 3 3668-QPYBK   Male             0      No         No      2          Yes
## 4 7795-CFOCW   Male             0      No         No     45           No
## 5 9237-HQITU Female             0      No         No      2          Yes
## 6 9305-CDSKC Female             0      No         No      8          Yes
## # ... with 14 more variables: MultipleLines <chr>, InternetService <chr>,
## #   OnlineSecurity <chr>, OnlineBackup <chr>, DeviceProtection <chr>,
## #   TechSupport <chr>, StreamingTV <chr>, StreamingMovies <chr>,
## #   Contract <chr>, PaperlessBilling <chr>, PaymentMethod <chr>,
## #   MonthlyCharges <dbl>, TotalCharges <dbl>, Churn <chr>
```

```
churn_clean <- churn_dat %>%
       mutate(SeniorCit = recode(SeniorCitizen, `1` = "Yes", `0` = "No"))

count(churn_clean, SeniorCit)
```

```
## # A tibble: 2 x 2
##   SeniorCit     n
##       <chr> <int>
## 1        No  5901
## 2       Yes  1142
```

```
quantile(churn_clean$tenure)
```

```
##   0%  25%  50%  75% 100%
##    0    9   29   55   72
```

```
churn_clean <- churn_clean %>%
        # create levels for tenure to categorize into different levels
        mutate(tenure_levels = cut(tenure, breaks = quantile(tenure)))
churn_clean
```

```
## # A tibble: 7,043 x 23
##    customerID gender SeniorCitizen Partner Dependents tenure PhoneService
##         <chr>  <chr>         <int>   <chr>      <chr>  <int>        <chr>
##  1 7590-VHVEG Female             0     Yes         No      1           No
##  2 5575-GNVDE   Male             0      No         No     34          Yes
##  3 3668-QPYBK   Male             0      No         No      2          Yes
##  4 7795-CFOCW   Male             0      No         No     45           No
##  5 9237-HQITU Female             0      No         No      2          Yes
##  6 9305-CDSKC Female             0      No         No      8          Yes
##  7 1452-KIOVK   Male             0      No        Yes     22          Yes
##  8 6713-OKOMC Female             0      No         No     10           No
##  9 7892-POOKP Female             0     Yes         No     28          Yes
## 10 6388-TABGU   Male             0      No        Yes     62          Yes
## # ... with 7,033 more rows, and 16 more variables: MultipleLines <chr>,
## #   InternetService <chr>, OnlineSecurity <chr>, OnlineBackup <chr>,
## #   DeviceProtection <chr>, TechSupport <chr>, StreamingTV <chr>,
## #   StreamingMovies <chr>, Contract <chr>, PaperlessBilling <chr>,
## #   PaymentMethod <chr>, MonthlyCharges <dbl>, TotalCharges <dbl>,
## #   Churn <chr>, SeniorCit <chr>, tenure_levels <fctr>
```

Although I added a new variable `tenure_levels` with tenure divided into groups, I'm not following up with this variable without subject matter knowledge and not cleaning variable MultipleLines into values of "No" or "Yes" for the same reason. There is a need for more information in order to make decisions on the cleaning and variable transformations.

3.5. Answer questions

1. What is the percentage of churn?

```
print(dfSummary(churn_clean), file = "Customer_churn_analysis_summary.html")
```

```
## Output file written: Customer_churn_analysis_summary.html
```

Data has 7043 observations and 21 variables. From the function `dfSummary` we can see there are 11 missing values (less than 1%) in the "TotalCharges" variable. The percentage of churn corresponds to 26.5% (1869 out of 5174 cases).

```
head(churn_clean$TotalCharges)
```

```
## [1]   29.85 1889.50  108.15 1840.75  151.65  820.50
```

How do missing values distribute across data?

```
churn_mis <- churn_clean %>%
        mutate(Missing_TotalCharges = is.na(TotalCharges))

churn_mis %>%
        ggplot(mapping = aes(x = MonthlyCharges, y = ..density..)) +
```

Data Frame Summary

churn_clean

N: 7043

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	customerID [character]	1. 0002-ORFBO 2. 0003-MKNFE 3. 0004-TLHLJ 4. 0011-IGKFF 5. 0013-EXCHZ 6. 0013-MHZWF 7. 0013-SMEOE 8. 0014-BMAQU 9. 0015-UOCOJ 10. 0016-QLJIS [7033 others]	1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 7033 (70.3%)		7043 (100%)	0 (0%)
2	gender [character]	1. Female 2. Male	3488 (49.5%) 3555 (50.5%)		7043 (100%)	0 (0%)
3	SeniorCitizen [integer]	mean (sd) : 0.16 (0.37) min < med < max : 0 < 0 < 1 IQR (CV) : 0 (2.27)	2 distinct val.		7043 (100%)	0 (0%)
4	Partner [character]	1. No 2. Yes	3641 (51.7%) 3402 (48.3%)		7043 (100%)	0 (0%)
5	Dependents [character]	1. No 2. Yes	4933 (70.0%) 2110 (30.0%)		7043 (100%)	0 (0%)
6	tenure [integer]	mean (sd) : 32.37 (24.56) min < med < max : 0 < 29 < 72 IQR (CV) : 46 (0.76)	73 distinct val.		7043 (100%)	0 (0%)
7	PhoneService [character]	1. No 2. Yes	682 (9.7%) 6361 (90.3%)		7043 (100%)	0 (0%)
8	MultipleLines [character]	1. No 2. No phone service 3. Yes	3390 (48.1%) 682 (9.7%) 2971 (42.2%)		7043 (100%)	0 (0%)
9	InternetService [character]	1. DSL 2. Fiber optic 3. No	2421 (34.4%) 3096 (44.0%) 1526 (21.7%)		7043 (100%)	0 (0%)
10	OnlineSecurity [character]	1. No 2. No internet service 3. Yes	3498 (49.7%) 1526 (21.7%) 2019 (28.7%)		7043 (100%)	0 (0%)
11	OnlineBackup [character]	1. No 2. No internet service 3. Yes	3088 (43.8%) 1526 (21.7%) 2429 (34.5%)		7043 (100%)	0 (0%)
12	DeviceProtection [character]	1. No 2. No internet service 3. Yes	3095 (43.9%) 1526 (21.7%) 2422 (34.4%)		7043 (100%)	0 (0%)
13	TechSupport [character]	1. No 2. No internet service 3. Yes	3473 (49.3%) 1526 (21.7%) 2044 (29.0%)		7043 (100%)	0 (0%)
14	StreamingTV [character]	1. No 2. No internet service 3. Yes	2810 (39.9%) 1526 (21.7%) 2707 (38.4%)		7043 (100%)	0 (0%)
15	StreamingMovies [character]	1. No 2. No internet service 3. Yes	2785 (39.5%) 1526 (21.7%) 2732 (38.8%)		7043 (100%)	0 (0%)
16	Contract [character]	1. Month-to-month 2. One year 3. Two year	3875 (55.0%) 1473 (20.9%) 1695 (24.1%)		7043 (100%)	0 (0%)
17	PaperlessBilling [character]	1. No 2. Yes	2872 (40.8%) 4171 (59.2%)		7043 (100%)	0 (0%)
18	PaymentMethod [character]	1. Bank transfer (automatic) 2. Credit card (automatic) 3. Electronic check 4. Mailed check	1544 (21.9%) 1522 (21.6%) 2365 (33.6%) 1612 (22.9%)		7043 (100%)	0 (0%)
19	MonthlyCharges [numeric]	mean (sd) : 64.76 (30.09) min < med < max : 18.25 < 70.35 < 118.75 IQR (CV) : 54.35 (0.46)	1585 distinct val.		7043 (100%)	0 (0%)
20	TotalCharges [numeric]	mean (sd) : 2283.3 (2266.77) min < med < max : 18.8 < 1397.47 < 8684.8 IQR (CV) : 3393.29 (0.99)	6530 distinct val.		7032 (99.84%)	11 (0.16%)
21	Churn [character]	1. No 2. Yes	5174 (73.5%) 1869 (26.5%)		7043 (100%)	0 (0%)
22	SeniorCit [character]	1. No 2. Yes	5901 (83.8%) 1142 (16.2%)		7043 (100%)	0 (0%)
23	tenure_levels [factor]	1. (0,9] 2. (9,29] 3. (29,55] 4. (55,72]	1843 (26.2%) 1715 (24.4%) 1719 (24.4%) 1755 (25.0%)		7032 (99.84%)	11 (0.16%)

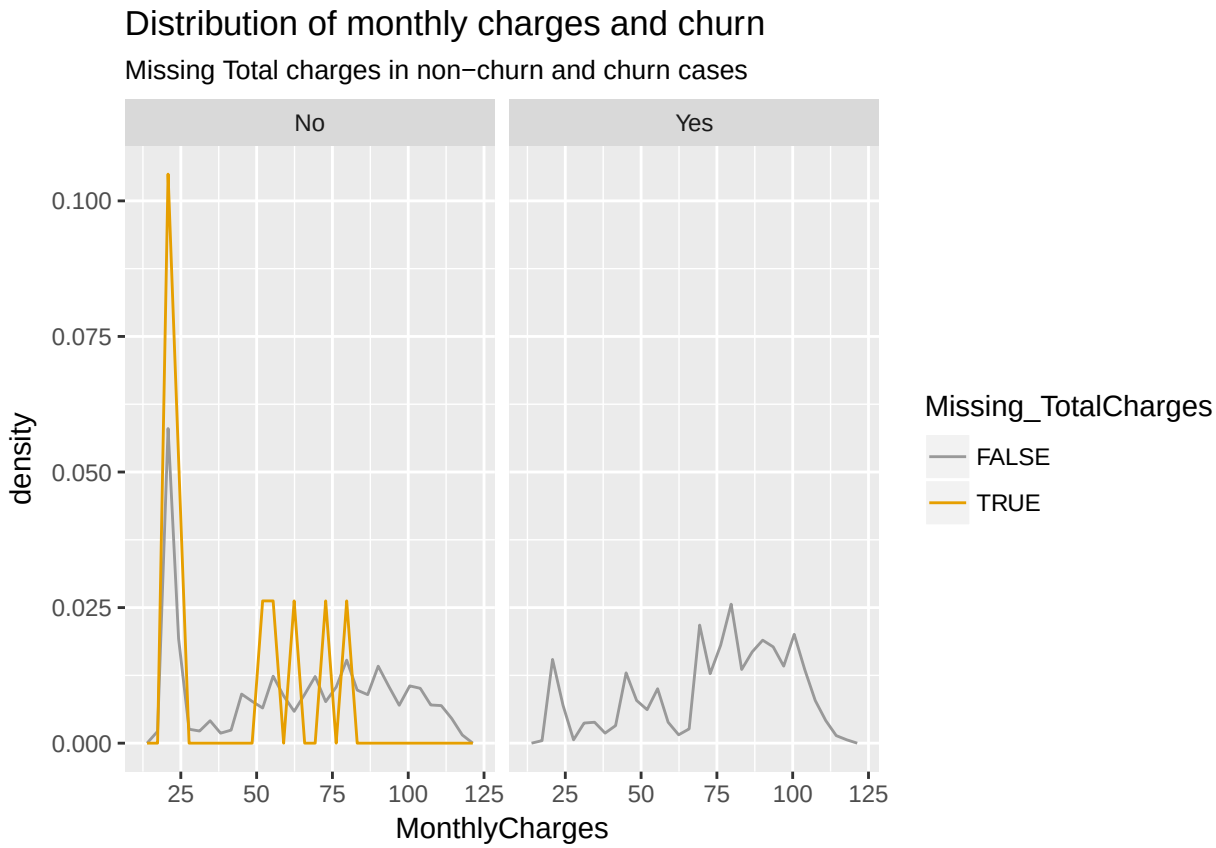
Generated by summarytools package version 0.8.1 (R version 3.4.1)
2018-02-01

Figure 1: Churn Summary Table

```
        geom_freqpoly(mapping = aes(colour = Missing_TotalCharges)) +
        facet_grid(~Churn) +
        scale_color_brewer(palette = "Set2") +
        ggtitle("Distribution of monthly charges and churn") +
        labs(subtitle = "Missing Total charges in non-churn and churn cases") +
        scale_color_manual(values=c("#999999", "#E69F00", "#56B4E9"))
```

```
## Scale for 'colour' is already present. Adding another scale for
## 'colour', which will replace the existing scale.
```

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

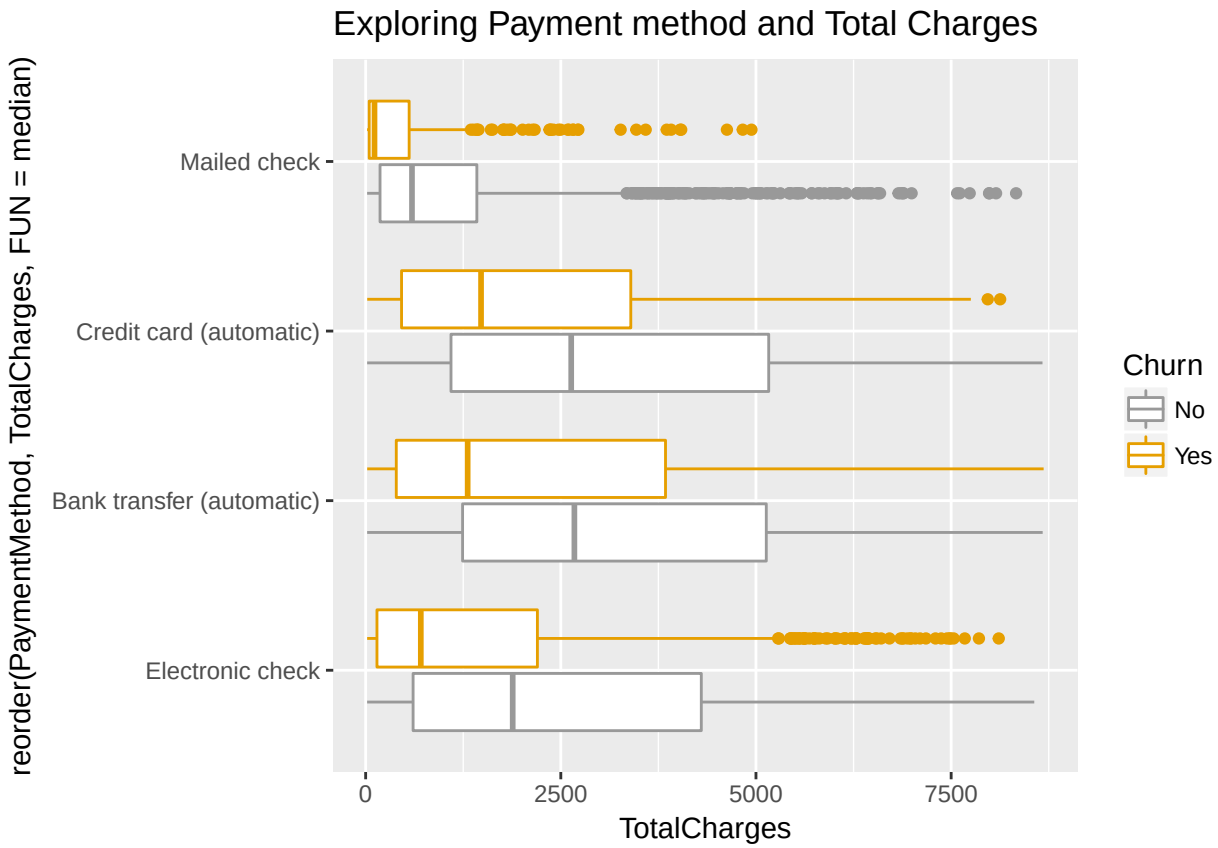


No missing values in churn cases, when we evaluate the proportional distribution of Monthly charges.

```
ggplot(data = churn_clean, mapping = aes(x = reorder(PaymentMethod, TotalCharges, FUN = median), y = To
        geom_boxplot() +
        scale_color_brewer(palette = "Set2") +
        coord_flip() +
        ggtitle("Exploring Payment method and Total Charges") +
        scale_color_manual(values=c("#999999", "#E69F00", "#56B4E9"))
```

```
## Scale for 'colour' is already present. Adding another scale for
## 'colour', which will replace the existing scale.
```

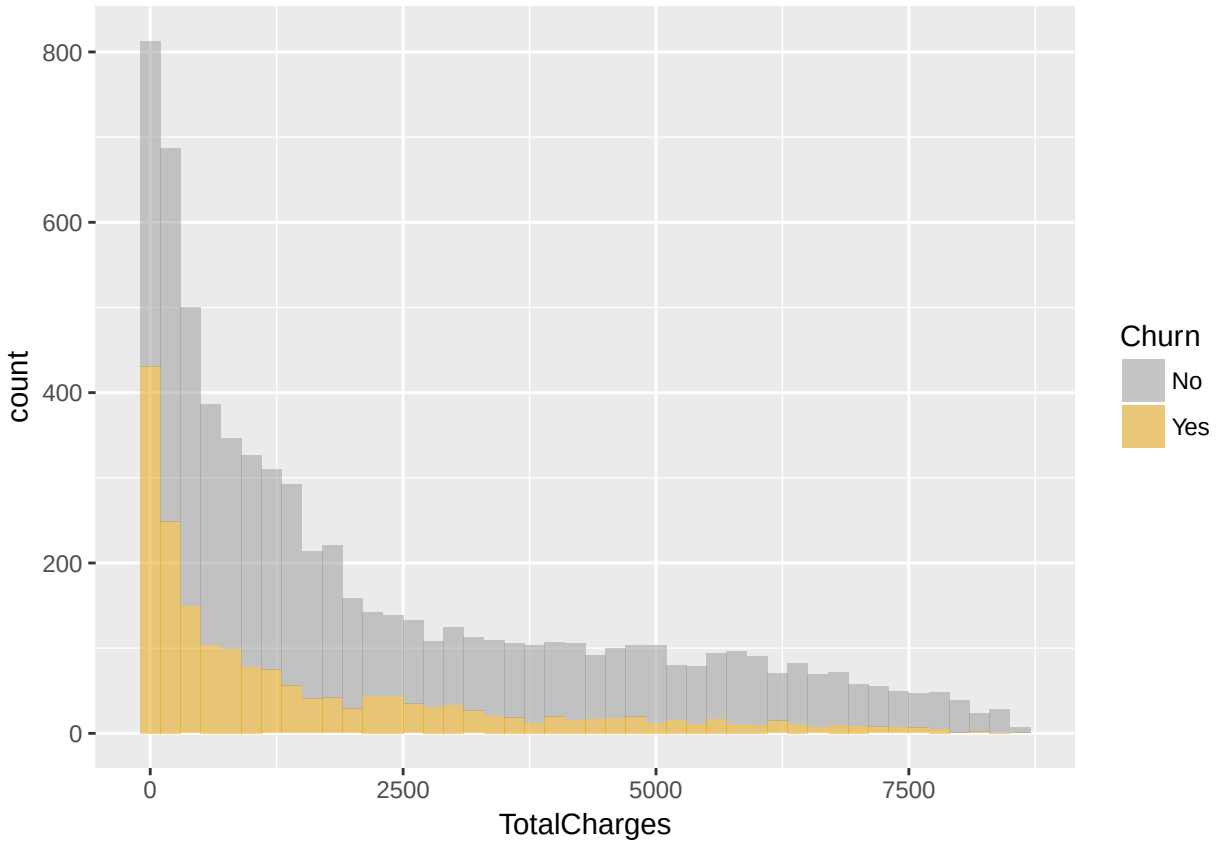
```
## Warning: Removed 11 rows containing non-finite values (stat_boxplot).
```



What is churn distribution?

```
ggplot(churn_clean) +
        geom_histogram(mapping = aes(x = TotalCharges, fill = Churn), alpha = 0.5, binwidth = 200) +
        scale_fill_manual(values=c("#999999", "#E69F00", "#56B4E9"))
```

```
## Warning: Removed 11 rows containing non-finite values (stat_bin).
```

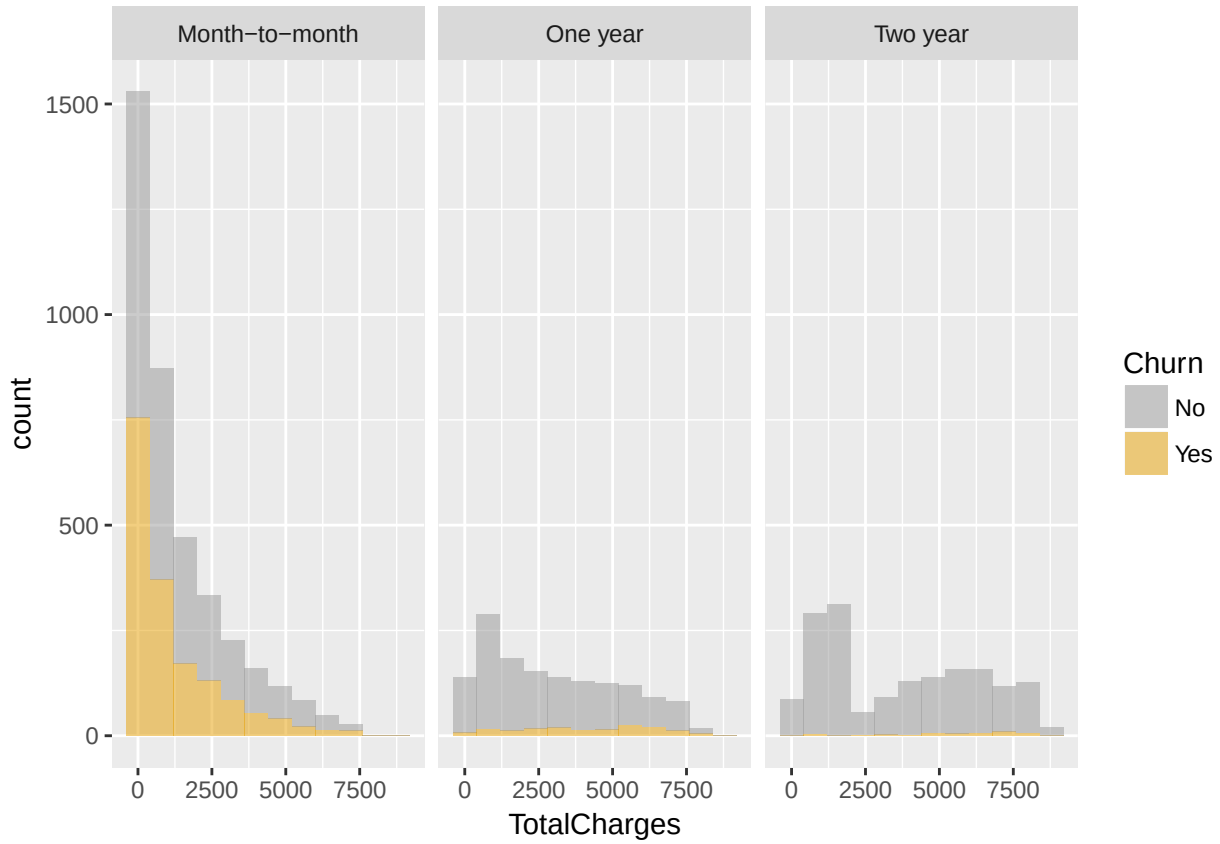


The density plot shows a skewed distribution for total charges in churn and non-churn cases. Here we can explore whether there's a tendency for lower total charges values in churn cases.

What is churn distribution per contract type/time?

```
ggplot(churn_clean) +
        geom_histogram(mapping = aes(x = TotalCharges, fill = Churn), alpha = 0.5, binwidth = 800) +
        facet_wrap(~Contract) +
        scale_fill_manual(values=c("#999999", "#E69F00", "#56B4E9"))
```

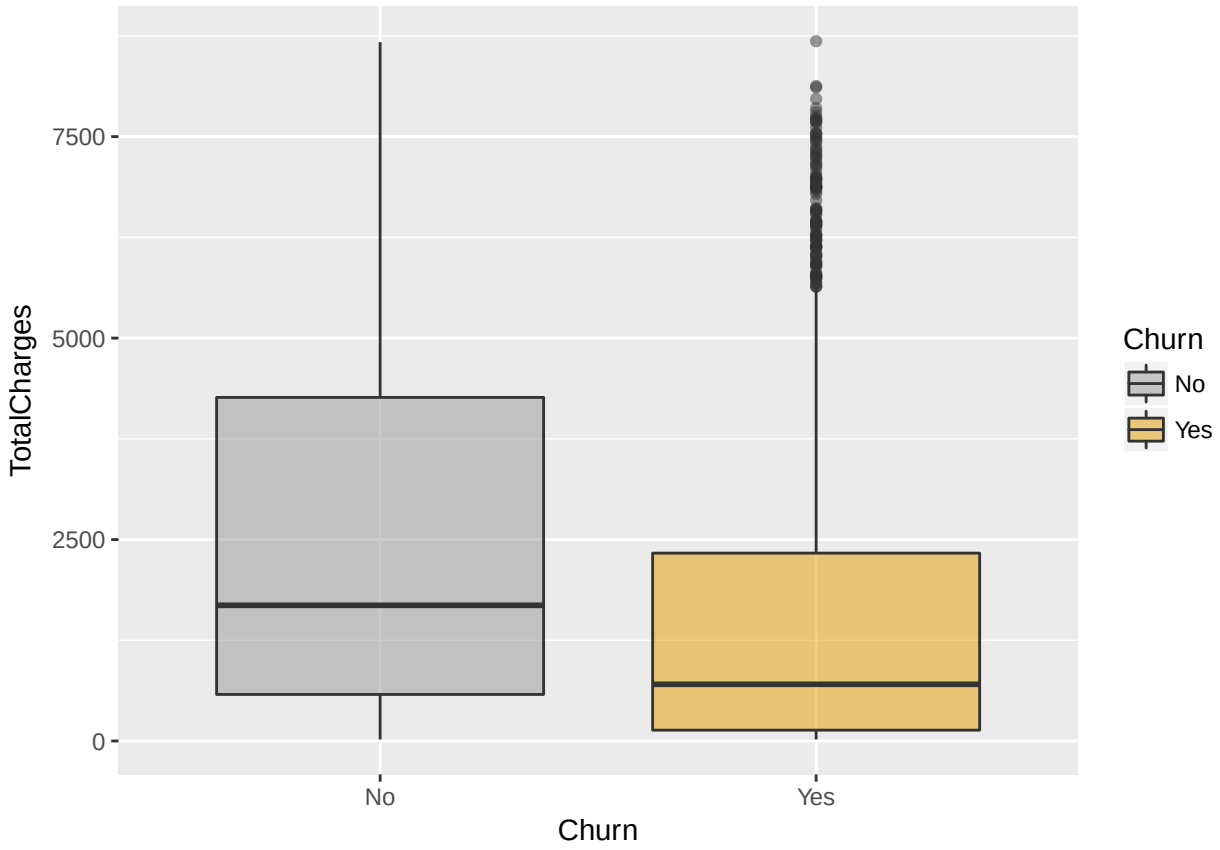
```
## Warning: Removed 11 rows containing non-finite values (stat_bin).
```




```
ggplot(churn_clean) +
        geom_boxplot(mapping = aes(x = Churn, y = TotalCharges, fill = Churn), alpha = 0.5, binwidth = 8
        scale_fill_manual(values=c("#999999", "#E69F00", "#56B4E9"))
```

```
## Warning: Ignoring unknown parameters: binwidth
```

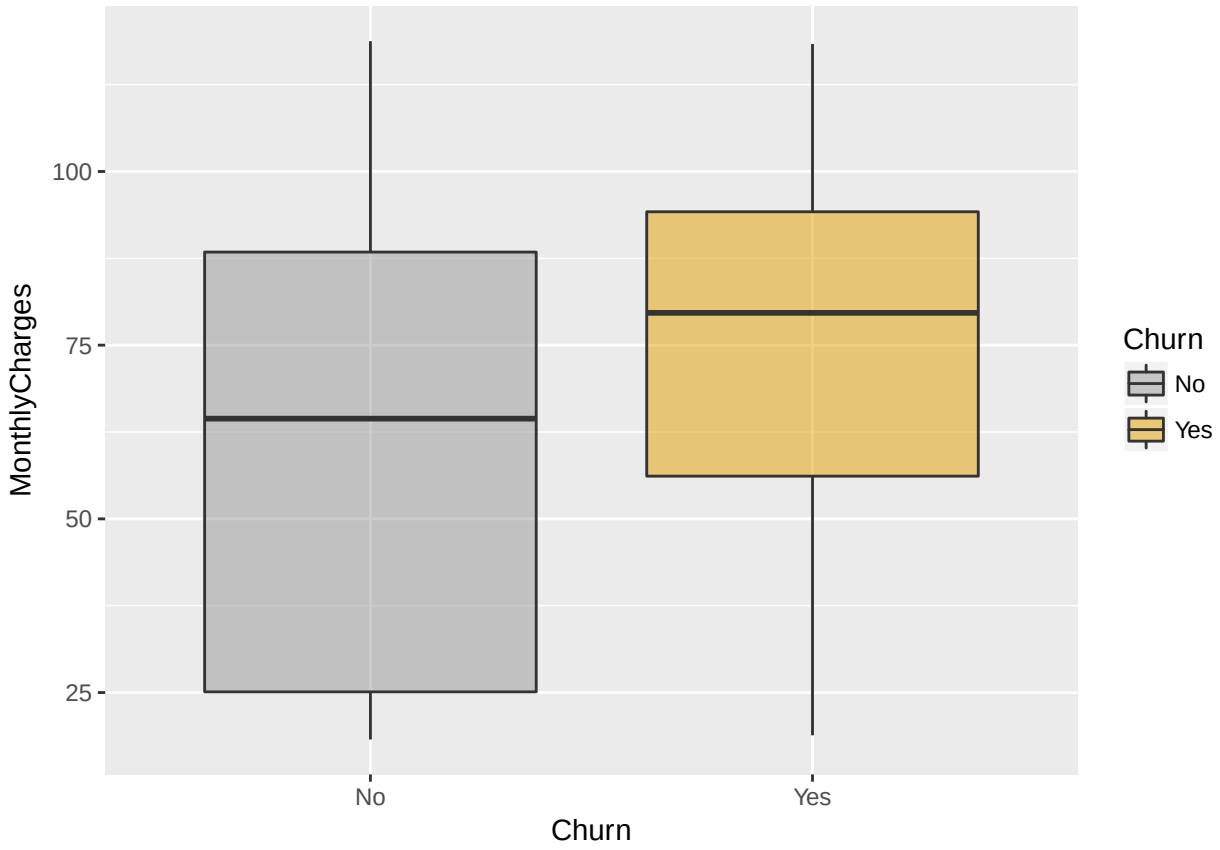
```
## Warning: Removed 11 rows containing non-finite values (stat_boxplot).
```



In average total charges are lower in churn cases than non-churn cases.

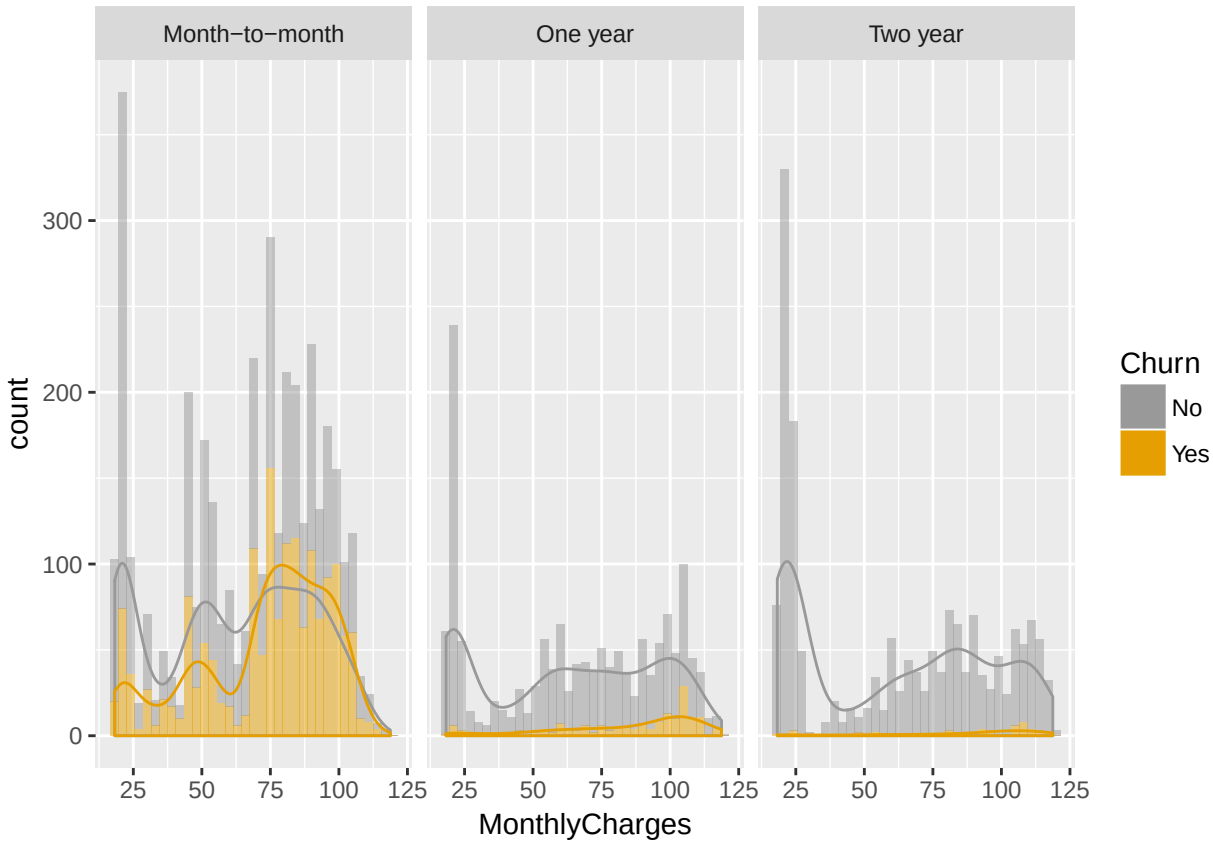
```
ggplot(churn_clean) +
        geom_boxplot(mapping = aes(x = Churn, y = MonthlyCharges, fill = Churn), alpha = 0.5, binwidth =
        scale_fill_manual(values=c("#999999", "#E69F00", "#56B4E9"))
```

```
## Warning: Ignoring unknown parameters: binwidth
```



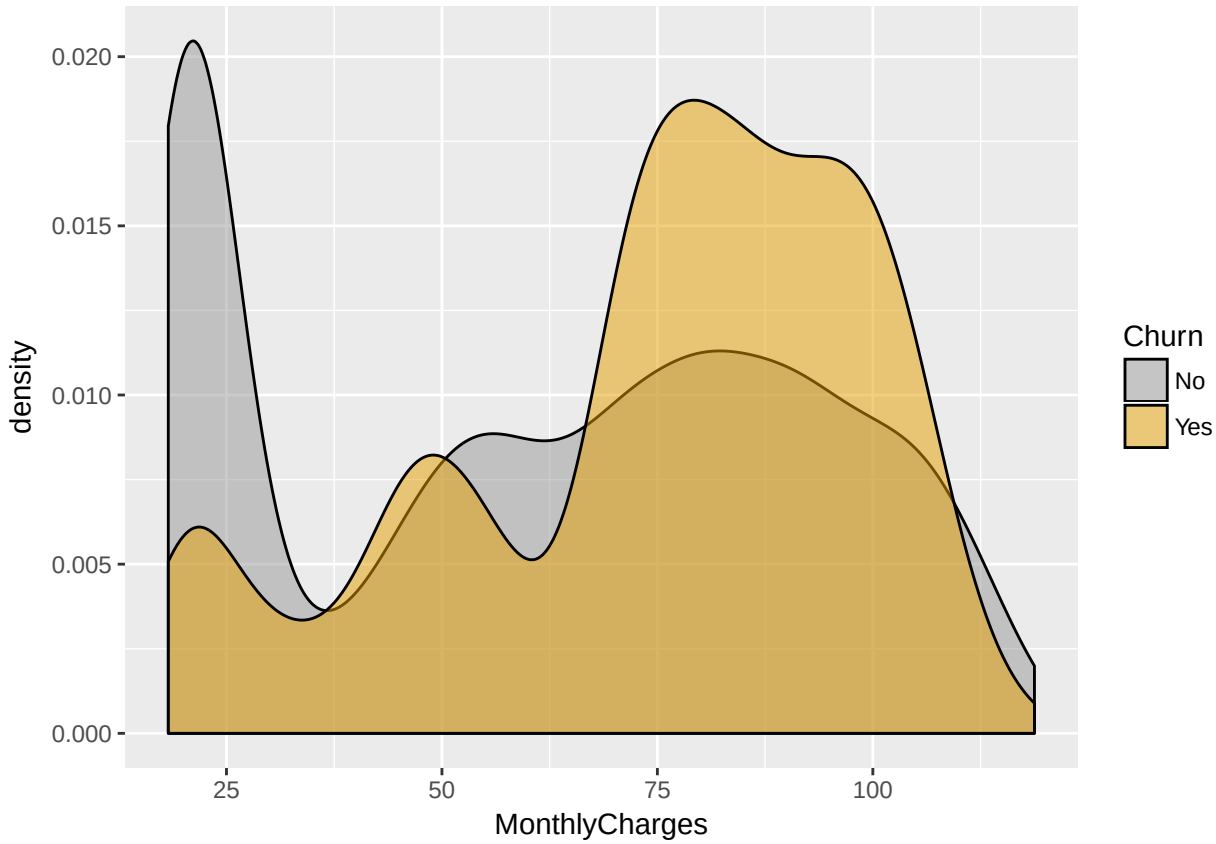
This plot contrast with the plot above showing churn cases with higher average of monthly costs. This deserves further exploration for monthly charges with a histogram.

```
ggplot(churn_mis) +
        geom_histogram(mapping = aes(x = MonthlyCharges, fill = Churn), alpha = 0.5, binwidth = 3) +
        geom_density(aes(y = 3 * ..count.., x = MonthlyCharges, color = Churn)) +
        facet_wrap(~Contract) +
        scale_fill_manual(values=c("#999999", "#E69F00", "#56B4E9")) +
        scale_color_manual(values=c("#999999", "#E69F00", "#56B4E9"))
```



This plot shows a different distribution of monthly charges per type of contract, showing more cases of churn in a month-to-month type of contract.

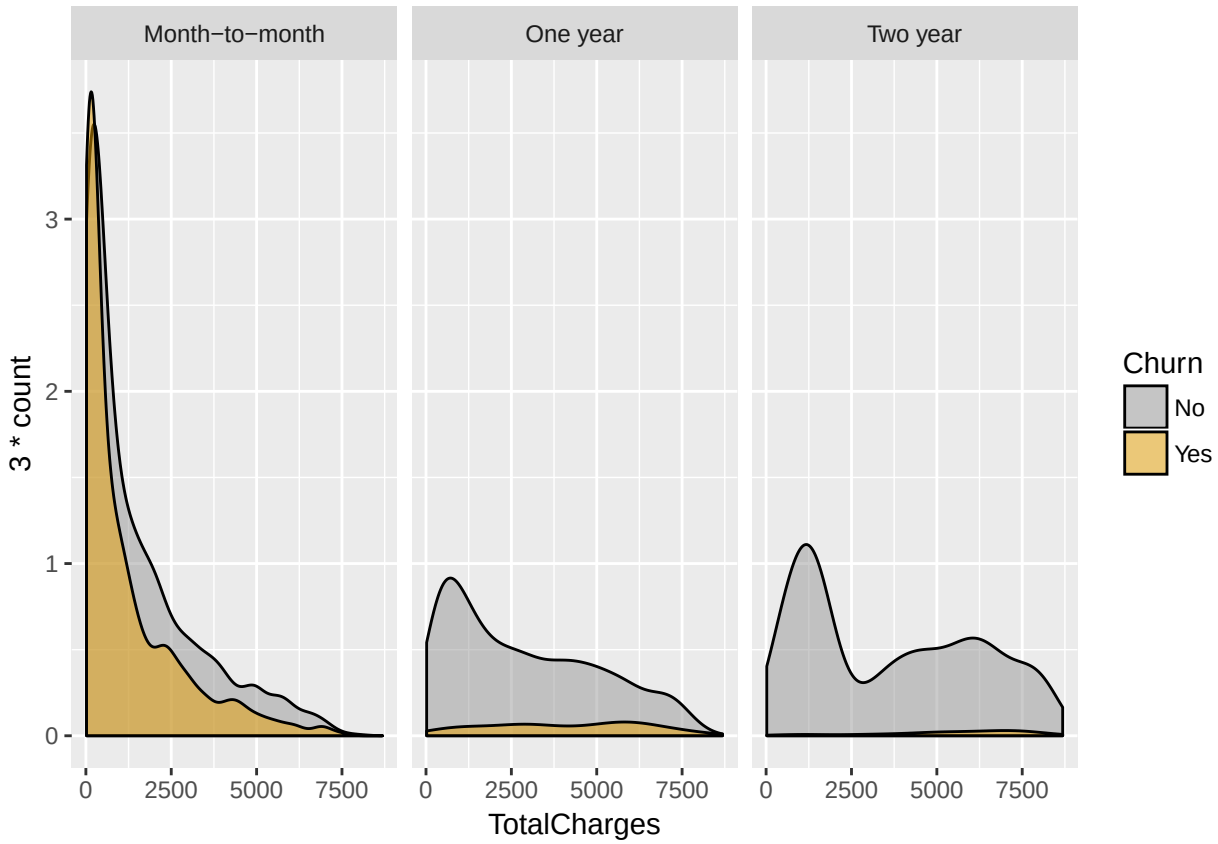
```
ggplot(churn_mis) +
        geom_density(mapping = aes(x = MonthlyCharges, fill = Churn), alpha = 0.5) +
        scale_fill_manual(values=c("#999999", "#E69F00", "#56B4E9"))
```



Proportionally there's more cases of churn with higher monthly charges.

```
ggplot(churn_mis) +
        geom_density(aes(y = 3 * ..count.., x = TotalCharges, fill = Churn), alpha = 0.5) +
 facet_grid(~Contract) +
        scale_fill_manual(values=c("#999999", "#E69F00", "#56B4E9"))
```

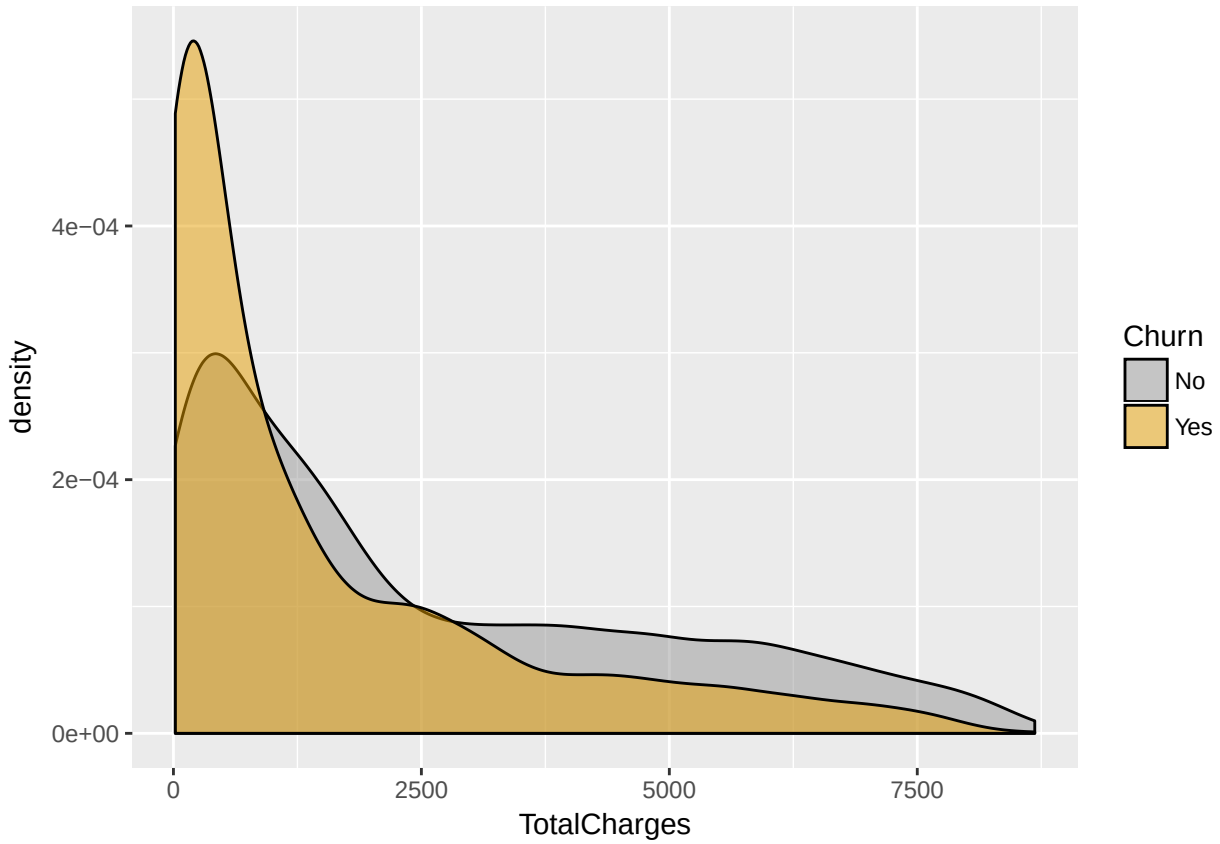
```
## Warning: Removed 11 rows containing non-finite values (stat_density).
```



We can see the same proportion for total charges distribution. With higher proportion of churn cases in a month-to-month type of contract.

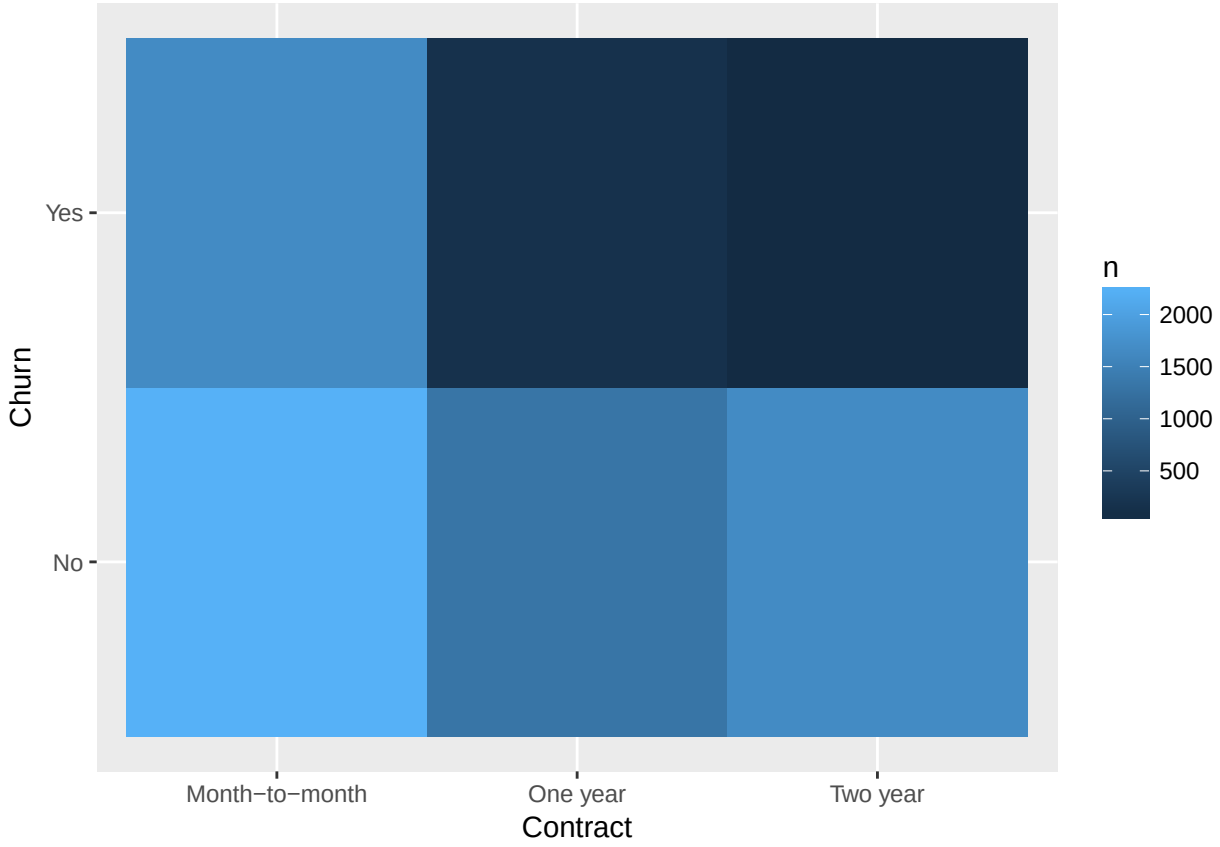
```
ggplot(churn_mis) +
        geom_density(mapping = aes(x = TotalCharges, fill = Churn), alpha = 0.5) +
        scale_fill_manual(values=c("#999999", "#E69F00", "#56B4E9"))
```

```
## Warning: Removed 11 rows containing non-finite values (stat_density).
```



Crossing two categorical variables (Churn and Contract) to understand where more number of Churn cases occur.

```
churn_mis %>%
        count(Churn, Contract) %>%
        ggplot(mapping = aes(x = Contract, y = Churn)) +
        geom_tile(mapping = aes(fill = n))
```



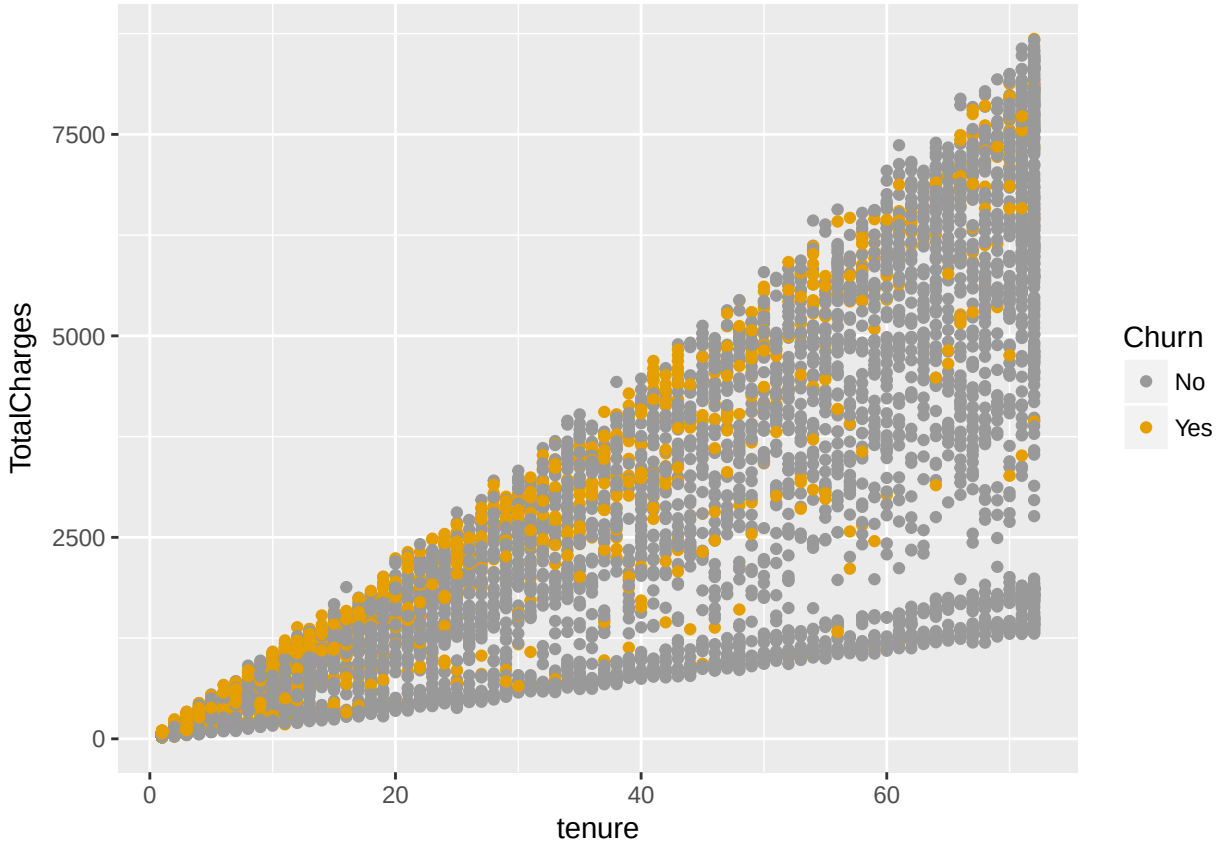
Removing missing values from the dataset

```
churn_new <- churn_mis %>%
        filter(complete.cases(.))
churn_new
```

```
## # A tibble: 7,032 x 24
##    customerID gender SeniorCitizen Partner Dependents tenure PhoneService
##         <chr>  <chr>         <int>   <chr>      <chr>  <int>        <chr>
##  1 7590-VHVEG Female             0     Yes         No      1           No
##  2 5575-GNVDE   Male             0      No         No     34          Yes
##  3 3668-QPYBK   Male             0      No         No      2          Yes
##  4 7795-CFOCW   Male             0      No         No     45           No
##  5 9237-HQITU Female             0      No         No      2          Yes
##  6 9305-CDSKC Female             0      No         No      8          Yes
##  7 1452-KIOVK   Male             0      No        Yes     22          Yes
##  8 6713-OKOMC Female             0      No         No     10           No
##  9 7892-POOKP Female             0     Yes         No     28          Yes
## 10 6388-TABGU   Male             0      No        Yes     62          Yes
## # ... with 7,022 more rows, and 17 more variables: MultipleLines <chr>,
## #   InternetService <chr>, OnlineSecurity <chr>, OnlineBackup <chr>,
## #   DeviceProtection <chr>, TechSupport <chr>, StreamingTV <chr>,
## #   StreamingMovies <chr>, Contract <chr>, PaperlessBilling <chr>,
## #   PaymentMethod <chr>, MonthlyCharges <dbl>, TotalCharges <dbl>,
## #   Churn <chr>, SeniorCit <chr>, tenure_levels <fctr>,
## #   Missing_TotalCharges <lgl>
```

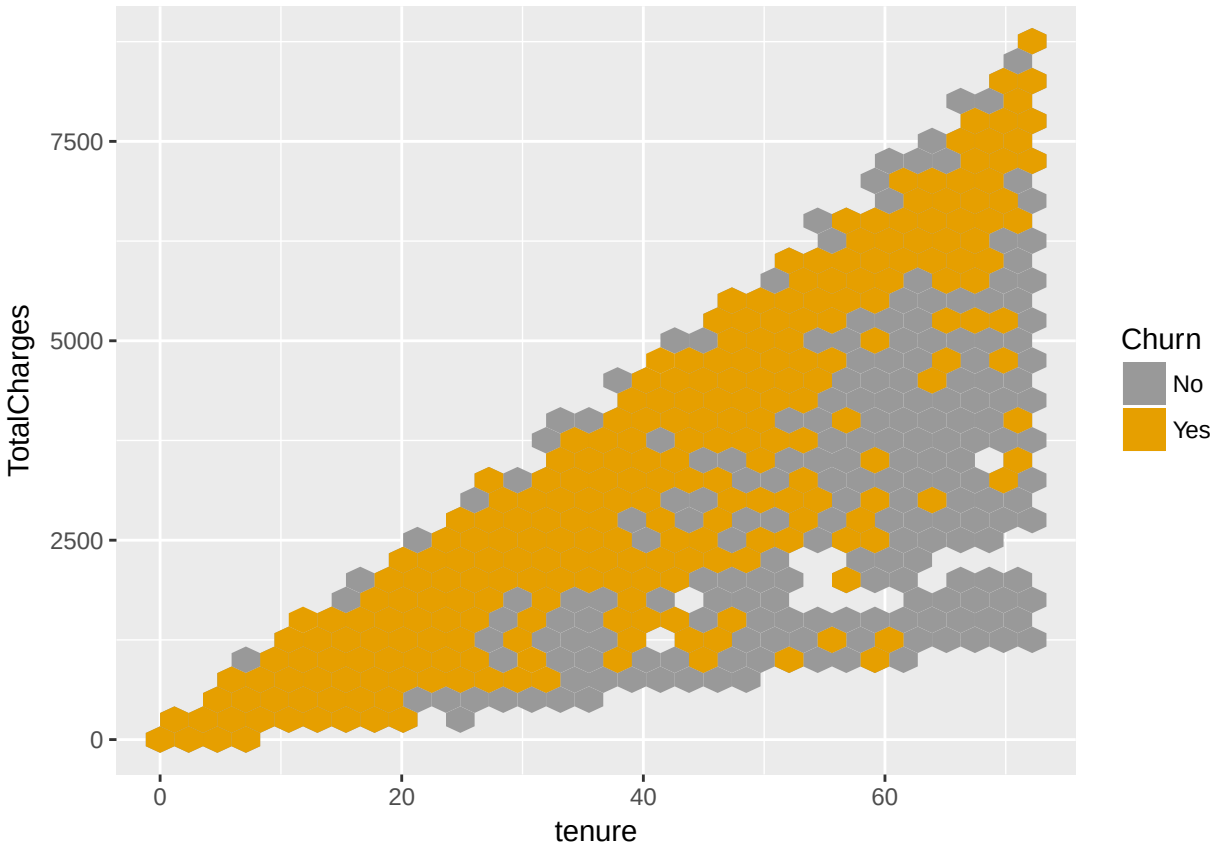
What are the variables associated to churn?

```
ggplot(churn_new) + geom_point(aes(y = TotalCharges, x = tenure, color = Churn)) +
        scale_color_manual(values=c("#999999", "#E69F00", "#56B4E9"))
```



Since we have many data points we'll use two dimentions bins (geom_hex() divides the coordinate plane into 2d bins)

```
ggplot(churn_new) +
        geom_hex(aes(y = TotalCharges, x = tenure, fill = Churn)) +
        scale_fill_manual(values=c("#999999", "#E69F00", "#56B4E9"))
```



This was an exploratory data analysis of the data. Further analysis to predict customer churn would require more knowledge about the data (such as the way data was collected and how each variable was defined and classified).

4. References

- https://datascienceplus.com/predict-customer-churn-logistic-regression-decision-tree-and-random-forest/
- http://r4ds.had.co.nz/
- http://dplyr.tidyverse.org/reference/recode.html