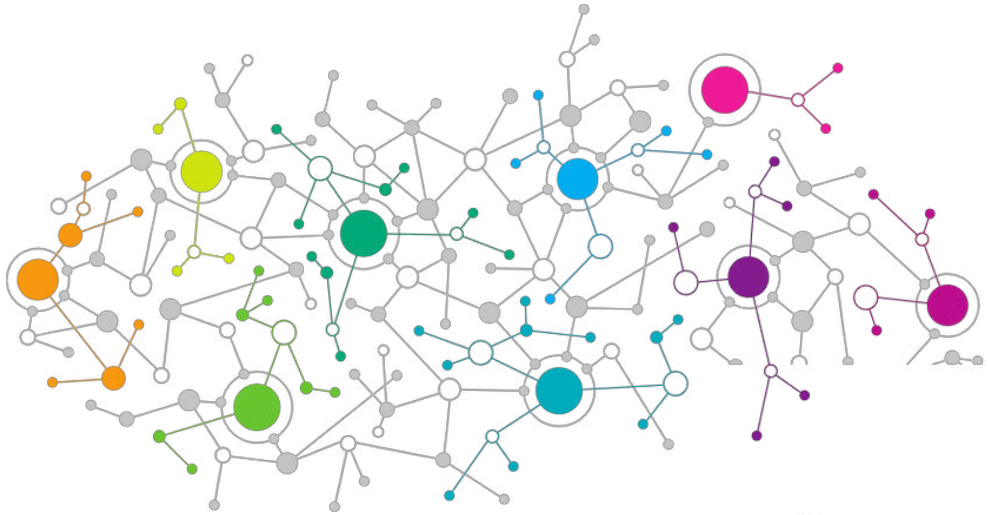
IntelliPaat

Data Science

Data Visualization



Agenda

01 Understanding Data Visualization

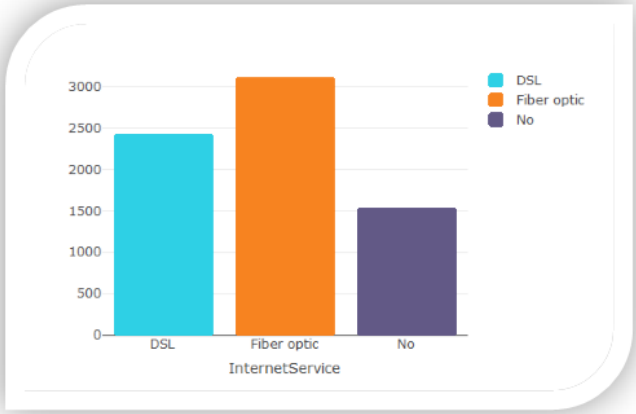
02 Base Graphics in R

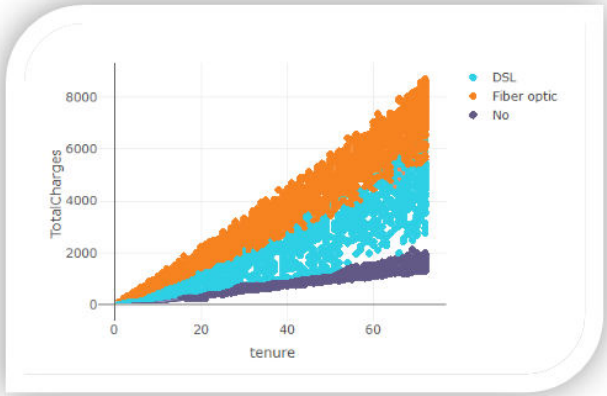
03 Visualization with GGPLOT2

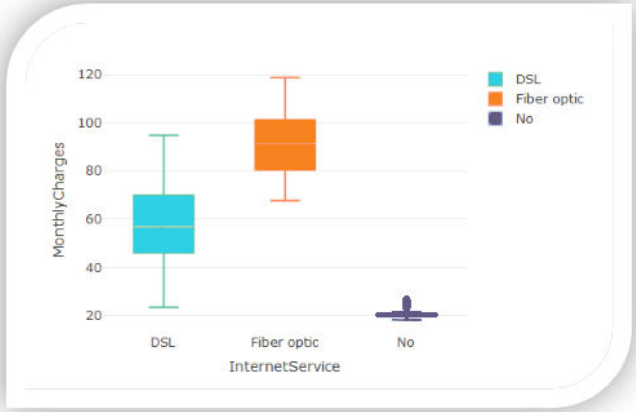
Data Visualization

Data Visualization

Data visualization is basically presentation of data with graphics. It's a way to summarize your findings and display it in a form that facilitates interpretation and can help in identifying patterns or trends







Importance of Data Visualization

Importance of Data Visualization

A picture is worth a 1000 words. Humans are easily attracted to visuals and most of us prefer to understand a particular scenario through pictures instead of text

It is faster to recognize a result than to read a paragraph. When done well, visualizations explain complex ideas simply. Oftentimes charts tell the story much faster than a prose description or a table presentation

Importance of Data Visualization

After looking at the data

Is it possible to answer the following questions?

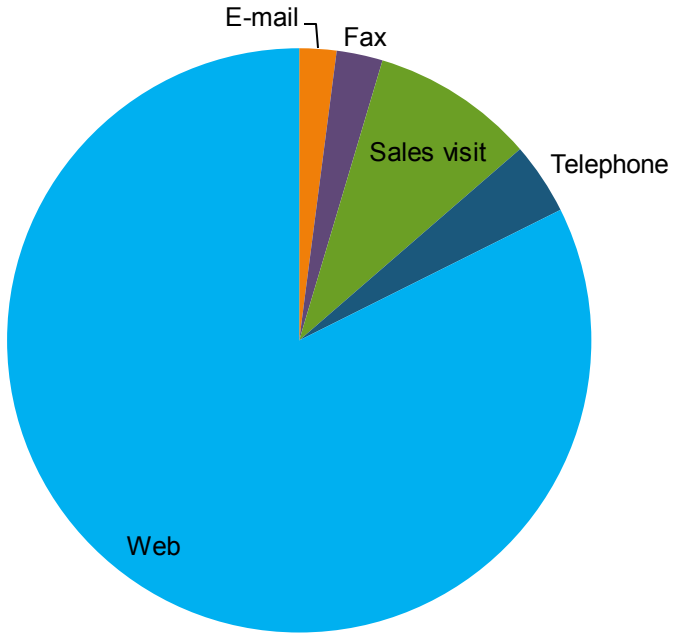
- **Which order method type has highest revenue ?**
- **Which product line has highest quantity ?**

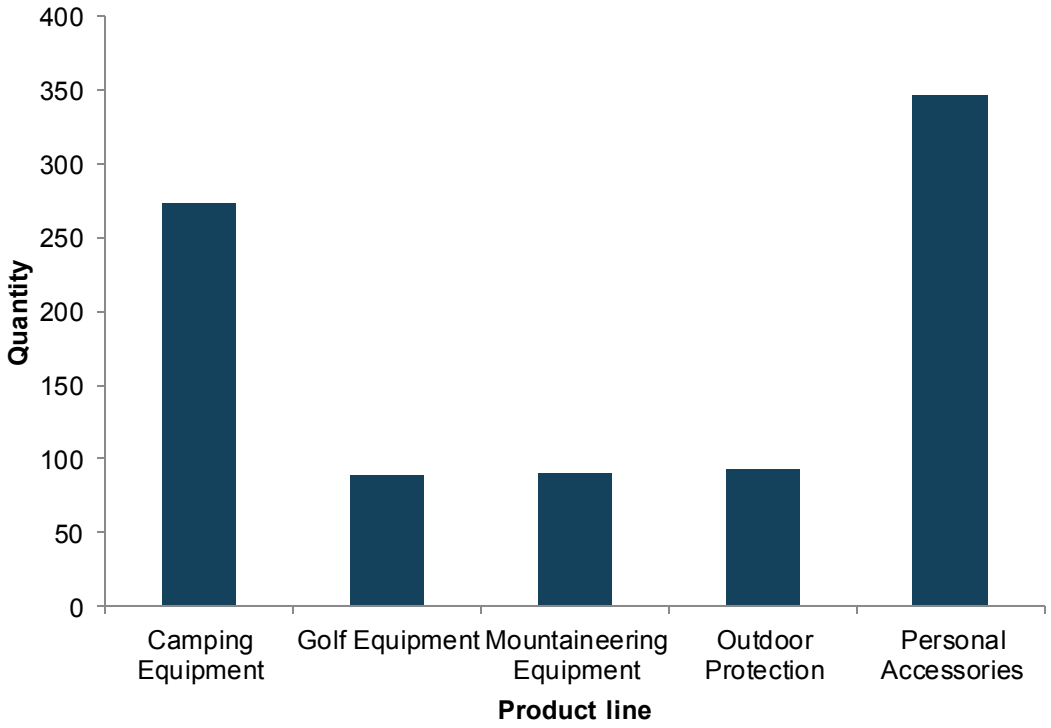
Below is the sample data: Sales Products

Retailer country	Order method type	Retailer type	Product line	Product type	Product	Year	Revenue	Quantity	Gross margin
United States	Telephone	Golf Shop	Personal Accessories	Navigation	Trail Master	2012	1095	3	0.34767123
United States	Telephone	Department Store	Camping Equipment	Sleeping Bags	Hibernator	2012	160103.2	1160	0.3769019
United States	Telephone	Department Store	Camping Equipment	Sleeping Bags	Hibernator Self - Inflating Mat	2012	66514.28	556	0.5440107
United States	Telephone	Department Store	Camping Equipment	Sleeping Bags	Hibernator Pad	2012	16205.73	411	0.51382196
United States	Telephone	Department Store	Camping Equipment	Sleeping Bags	Hibernator Pillow	2012	33520.42	2475	0.46298346
United States	Fax	Outdoors Shop	Camping Equipment	Cooking Gear	TrailChef Water Bag	2013	19418.52	3102	0.53194888
United States	Fax	Outdoors Shop	Camping Equipment	Cooking Gear	TrailChef Cook Set	2013	42304.32	794	0.34365616
United States	Fax	Outdoors Shop	Camping Equipment	Cooking Gear	TrailChef Single Flame	2013	52266.32	824	0.26880025
United States	Telephone	Department Store	Camping Equipment	Packs	Canyon Mule Journey Backpack	2012	235660.36	676	0.38805542
United States	Telephone	Department Store	Camping Equipment	Packs	Canyon Mule Cooler	2012	53822.16	1652	0.50890117
United States	Web	Golf Shop	Personal Accessories	Watches	Venue	2014	73949	1013	0.42858294
United States	Web	Golf Shop	Personal Accessories	Watches	Infinity	2014	157890.2	665	0.45986527
United States	Web	Golf Shop	Personal Accessories	Watches	Lux	2014	67265.2	396	0.48654044

Importance of Data Visualization

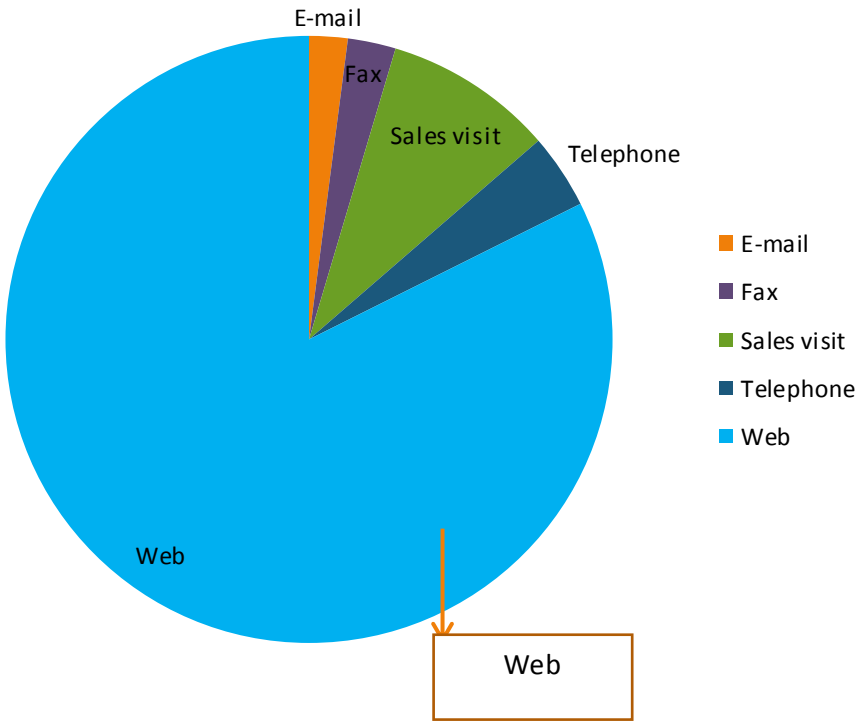
After displaying the data in the form of graphs . Lets try answering the questions





Importance of Data Visualization

After displaying the data in the form of graphs . Lets try answering the questions



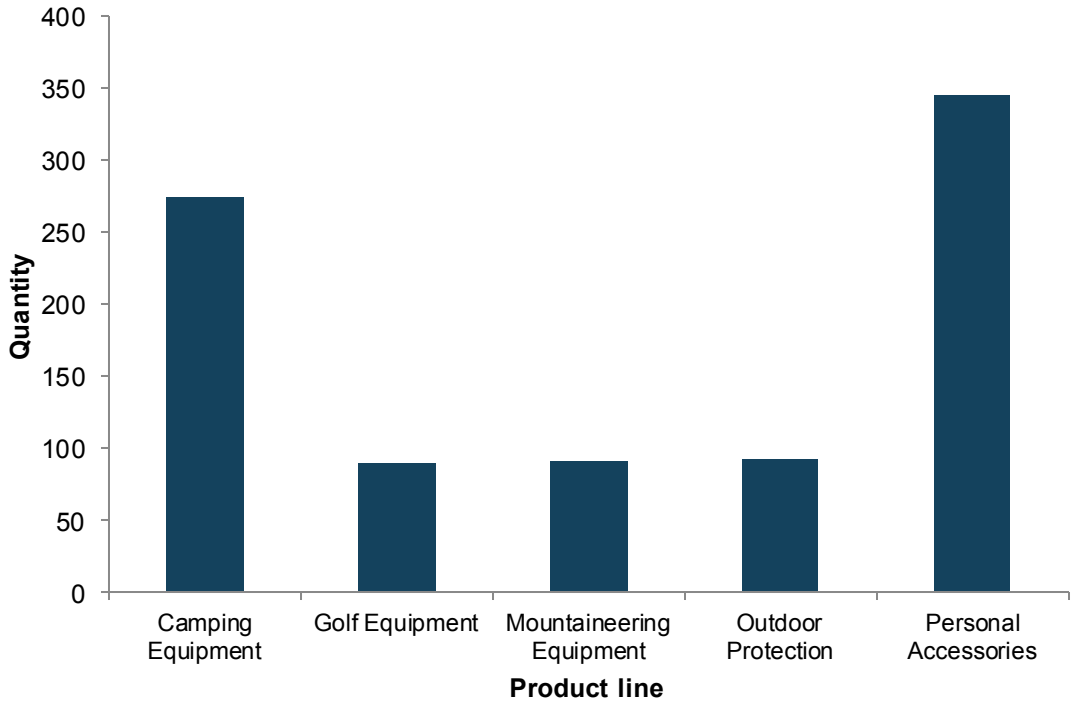
Which order method type has highest revenue ?

Web order method type has highest revenue among all the order method type.

Moreover looking at the pie chart you can also tell that which order method type has lowest revenue : Email.

Importance of Data Visualization

After displaying the data in the form of graphs . Lets try answering the questions



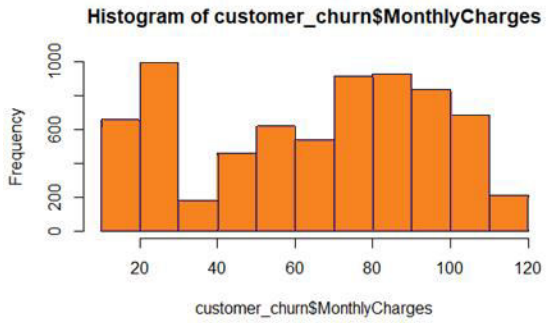
Which product line has highest quantity ?

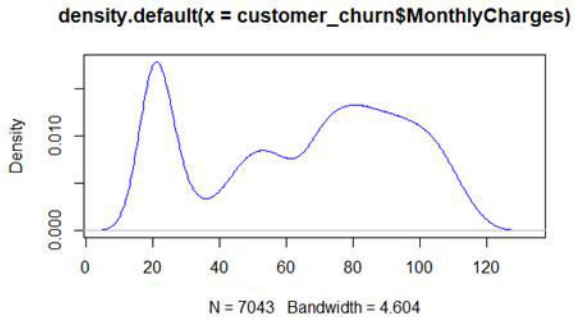
From the graph it can be easily seen that Personal Accessories has highest number among all the product line.

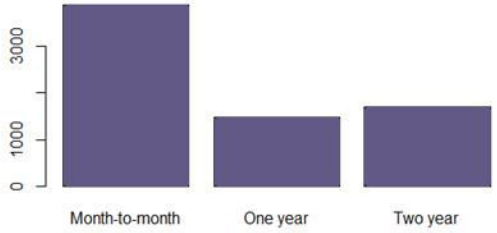
Base Graphics in R

Base Graphics in R

Base Graphics helps in making simple graphs



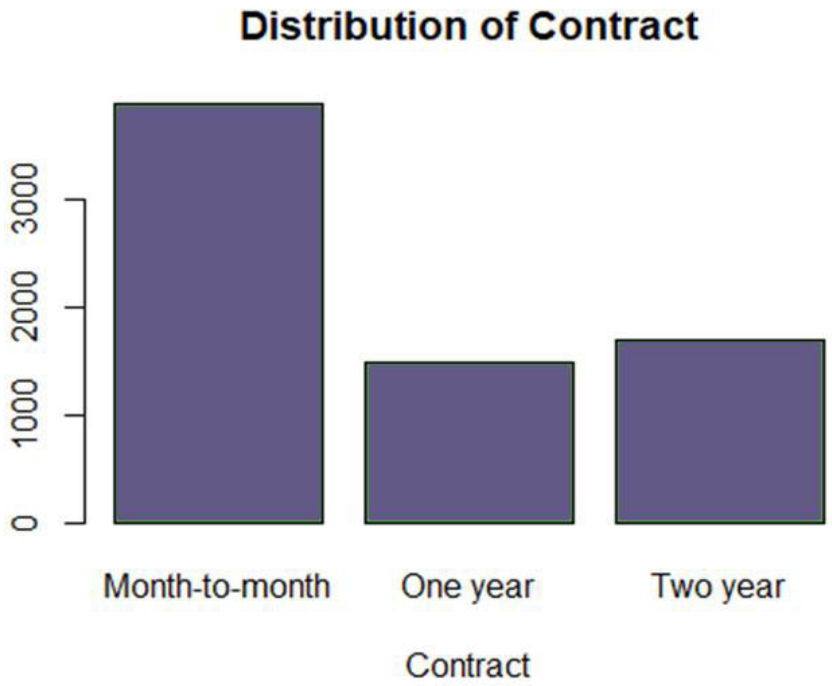




Bar Plot

Bar Plot

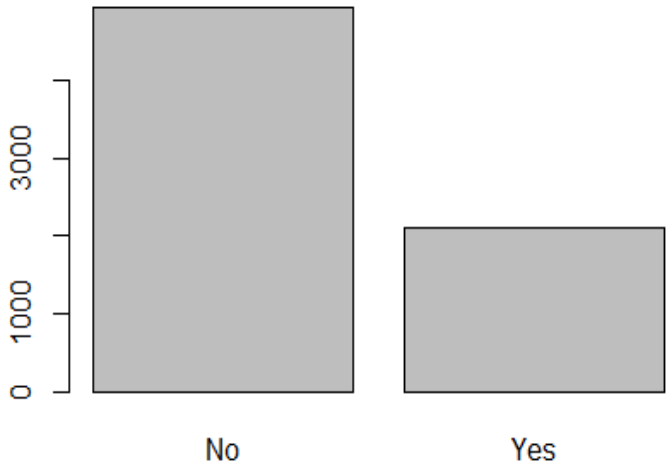
Bar Plots are suitable for showing comparison between cumulative totals across several groups



Bar Plot

Making a simple Bar-plot

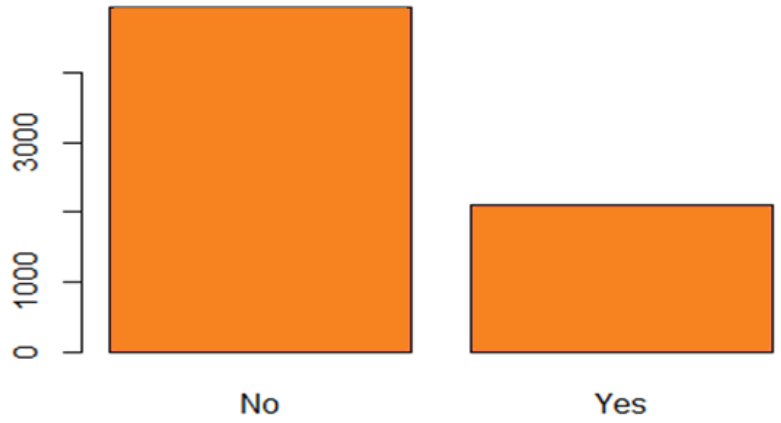
```
plot(customer_churn$Dependents)
```



Bar Plot

Adding color

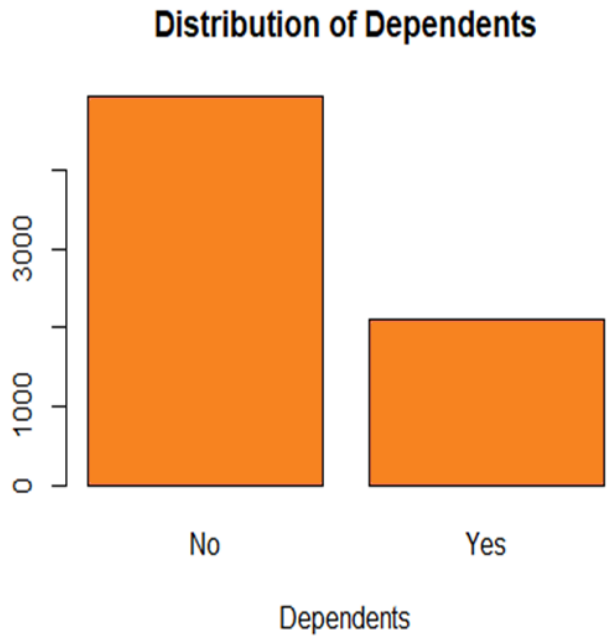
```
plot(customer_churn$Dependents, col="coral")
```



Bar Plot

Adding x-axis label & title

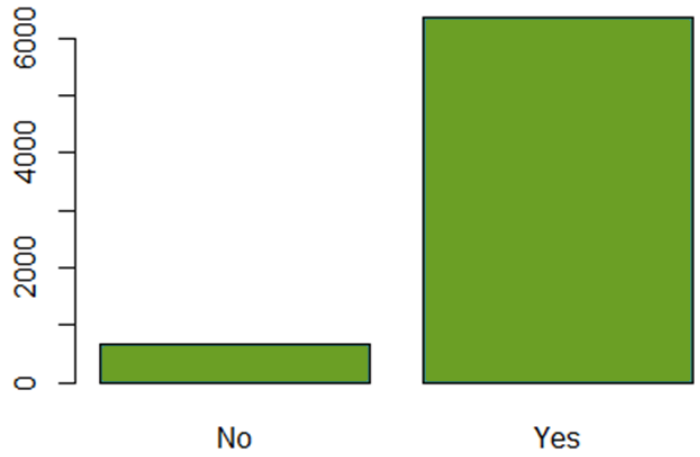
```
plot(customer_churn$Dependents,
col="coral",xlab="Dependents",
main="Distribution of Dependents")
```



Bar Plot

Bar-plot for ‘PhoneService’ column

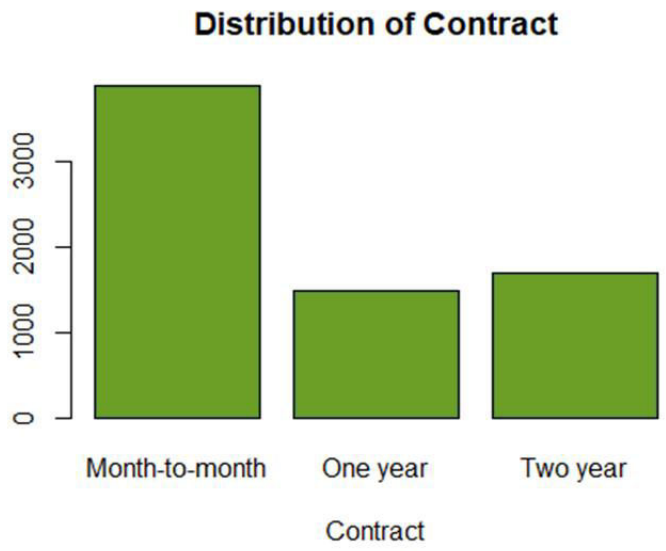
```
plot(customer_churn$PhoneService,col="aquamarine4"
)
```



Bar Plot

Bar-plot for ‘Contract’ column

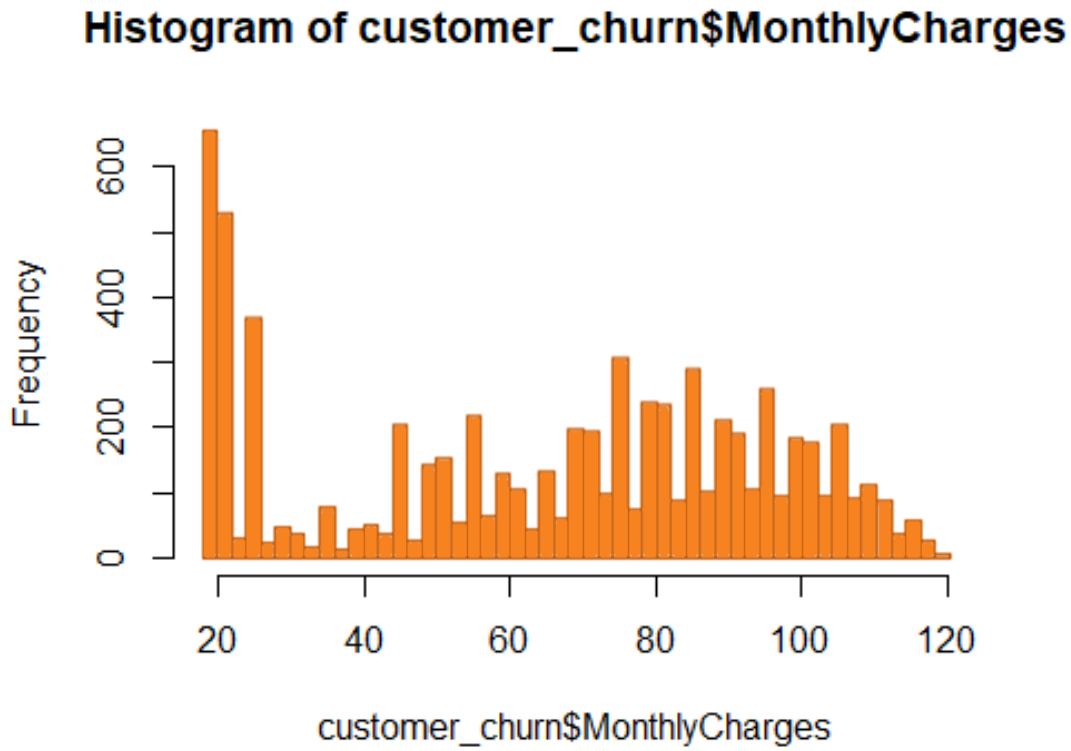
```
plot(customer_churn$Contract,col="palegreen4",
xlab="Contract",
main="Distribution of Contract")
```



Histogram

Histogram

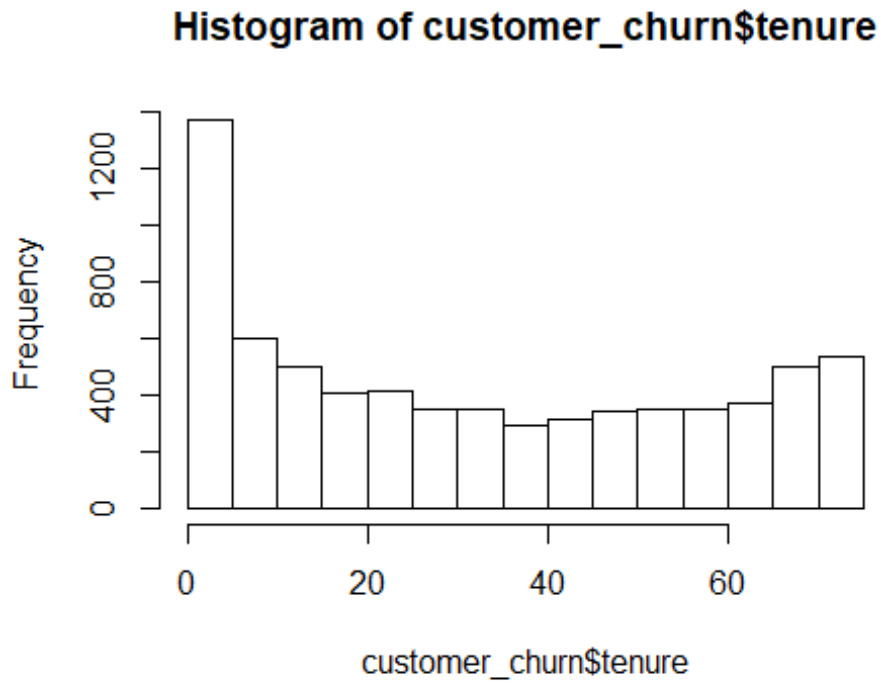
Histogram is basically a plot that breaks the data into bins (or breaks) and shows frequency distribution of these bins



Histogram

Making a simple histogram

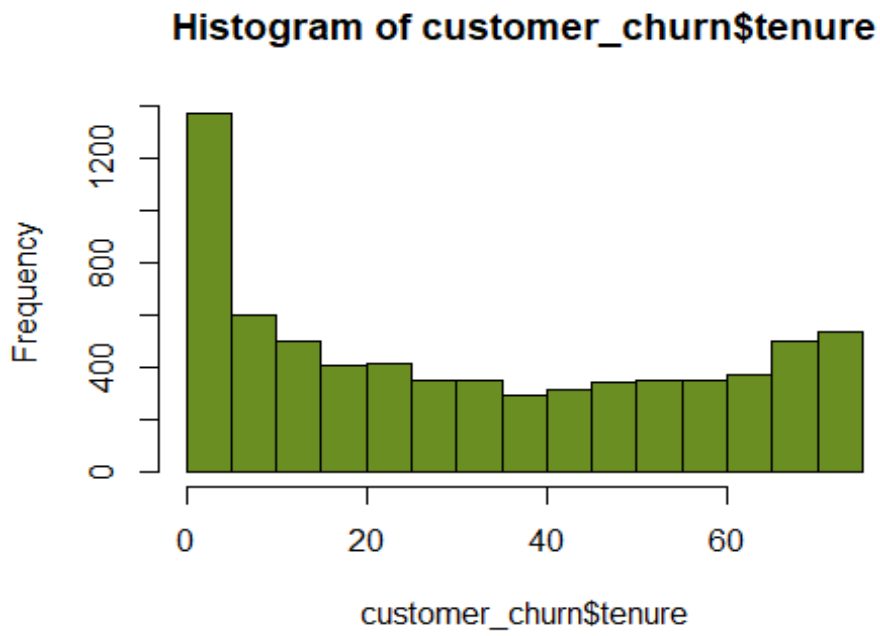
hist(customer_churn**$**tenure)



Histogram

Adding Color

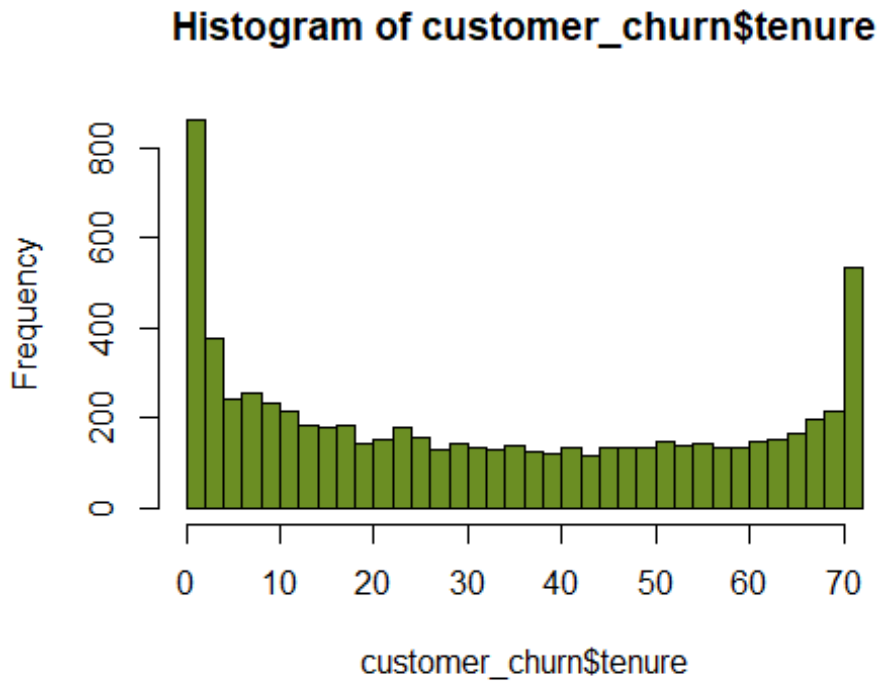
```
hist(customer_churn$tenure,col="olivedrab")
```



Histogram

Change the number of bins

```
hist(customer_churn$tenure,
col="olivedrab",
breaks=30)
```



Grammar of Graphics

Grammar of Graphics

Every form of communication needs to have grammar. Since, visualization is also a form of communication, it needs to have a foundation of grammar

I am John

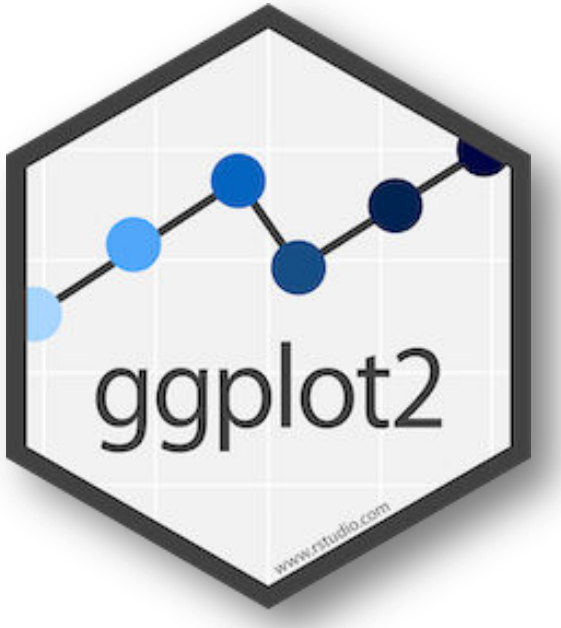
Am John I

Components of Grammar of Graphics

Element	Description
Data	The data-set for which we would want to plot a graph
Aesthetics	The metrics onto which we plot our data
Geometry	Visual elements to plot the data
Facet	Groups by which we divide the data

Visualization with ggplot2

Visualization with ggplot2

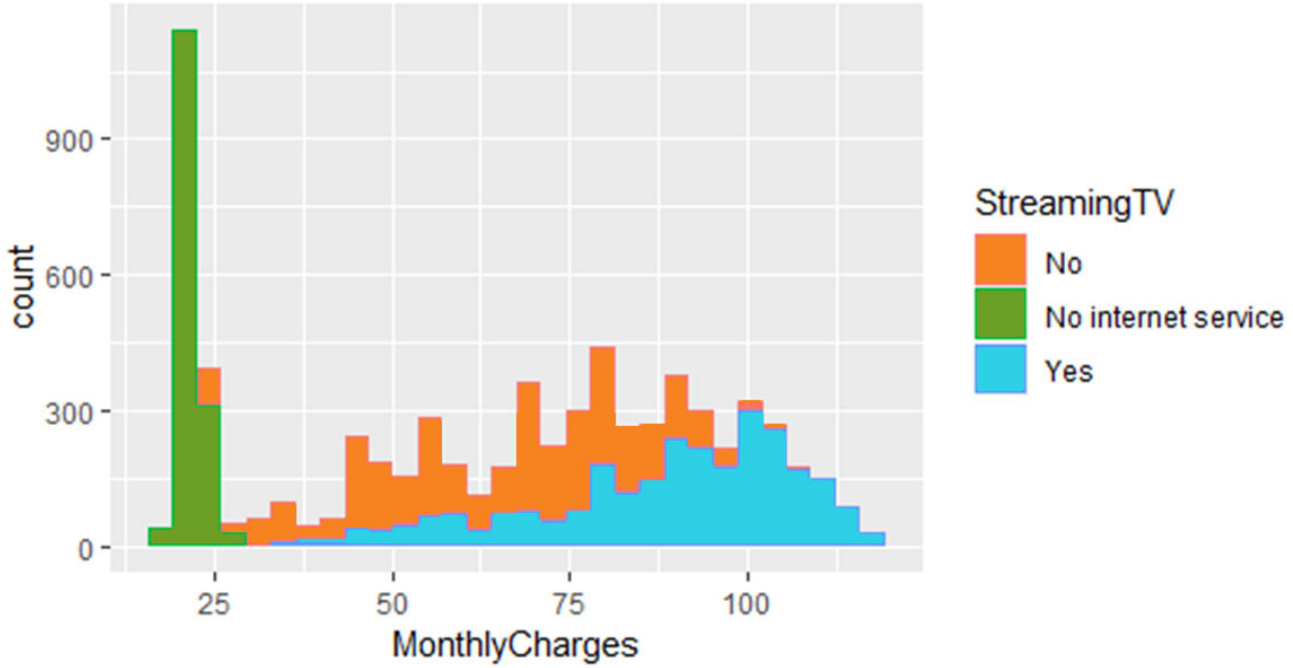


ggplot2 is a system for declaratively creating graphics, based on the Grammar of Graphics. You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details

geom_hist()

geom_hist()

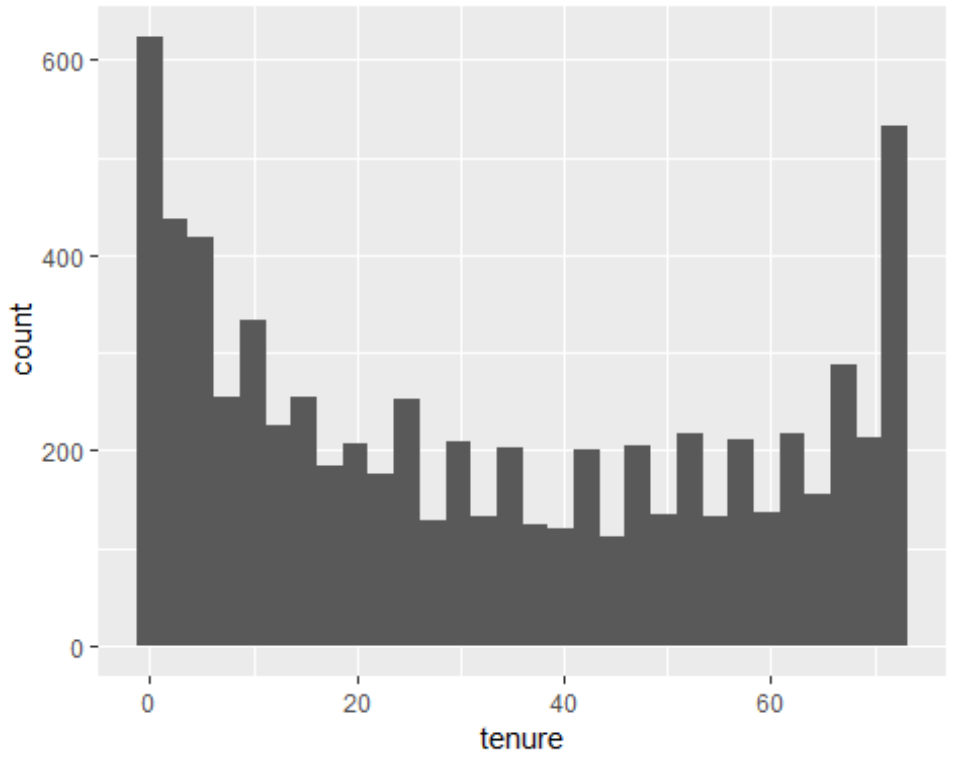
geom_hist() function helps in making histograms with ggplot2



geom_hist()

Build a histogram for ‘tenure’ column

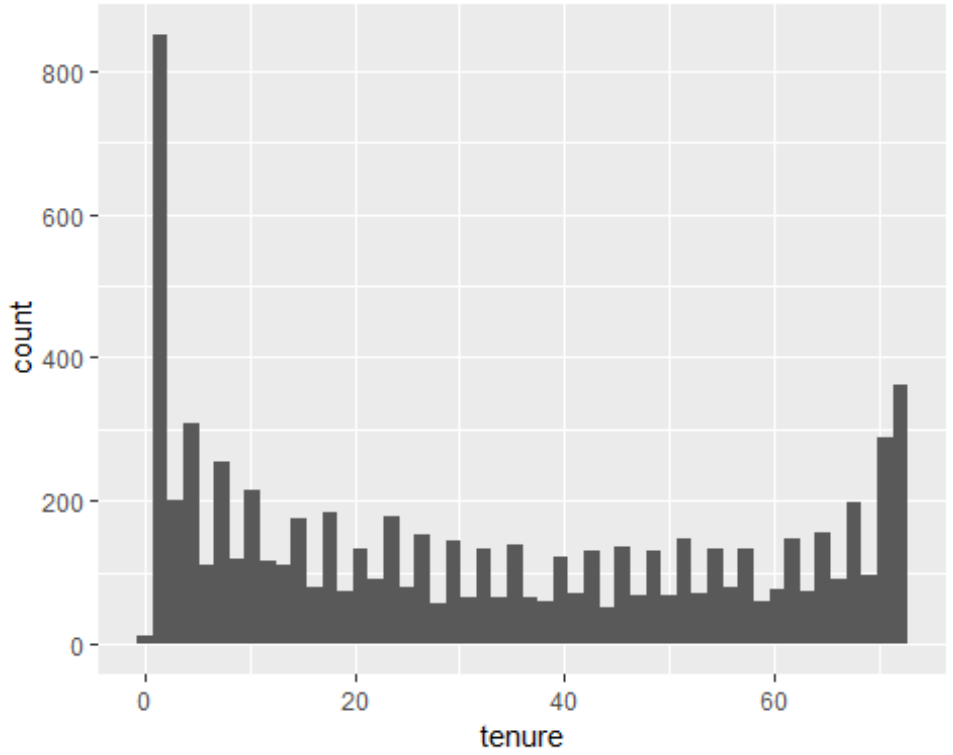
```
ggplot(data=customer_churn,
aes(x=tenure))+geom_histogram()
```



geom_hist()

Change the number of bins

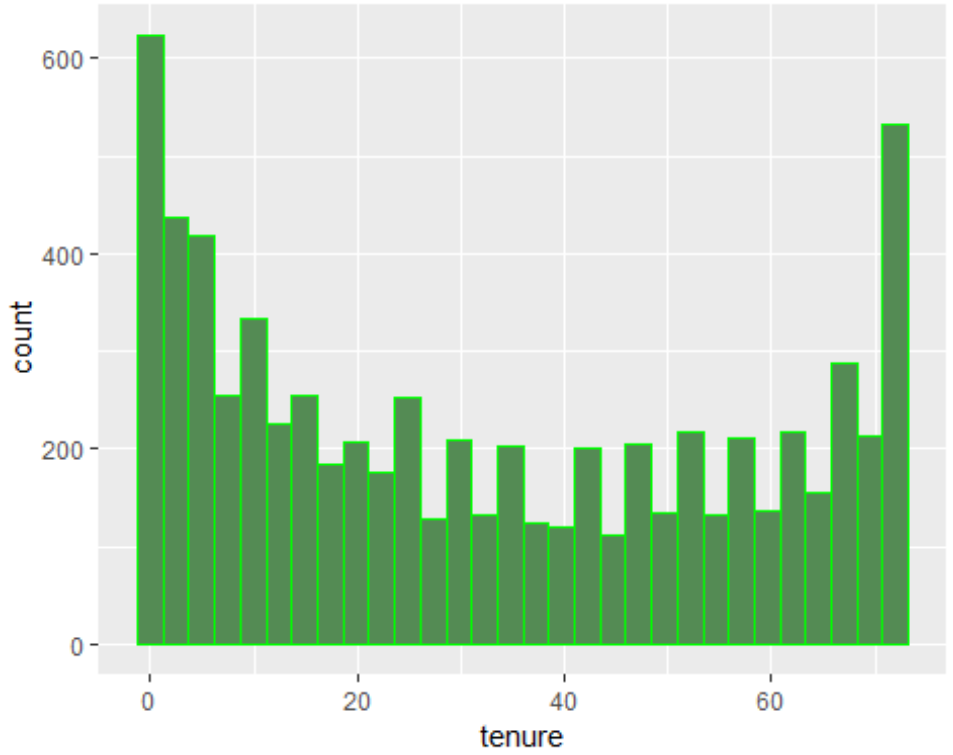
```
ggplot(data = customer_churn,
aes(x=tenure))+geom_histogram(bins = 50)
```



geom_hist()

Add fill color and border color

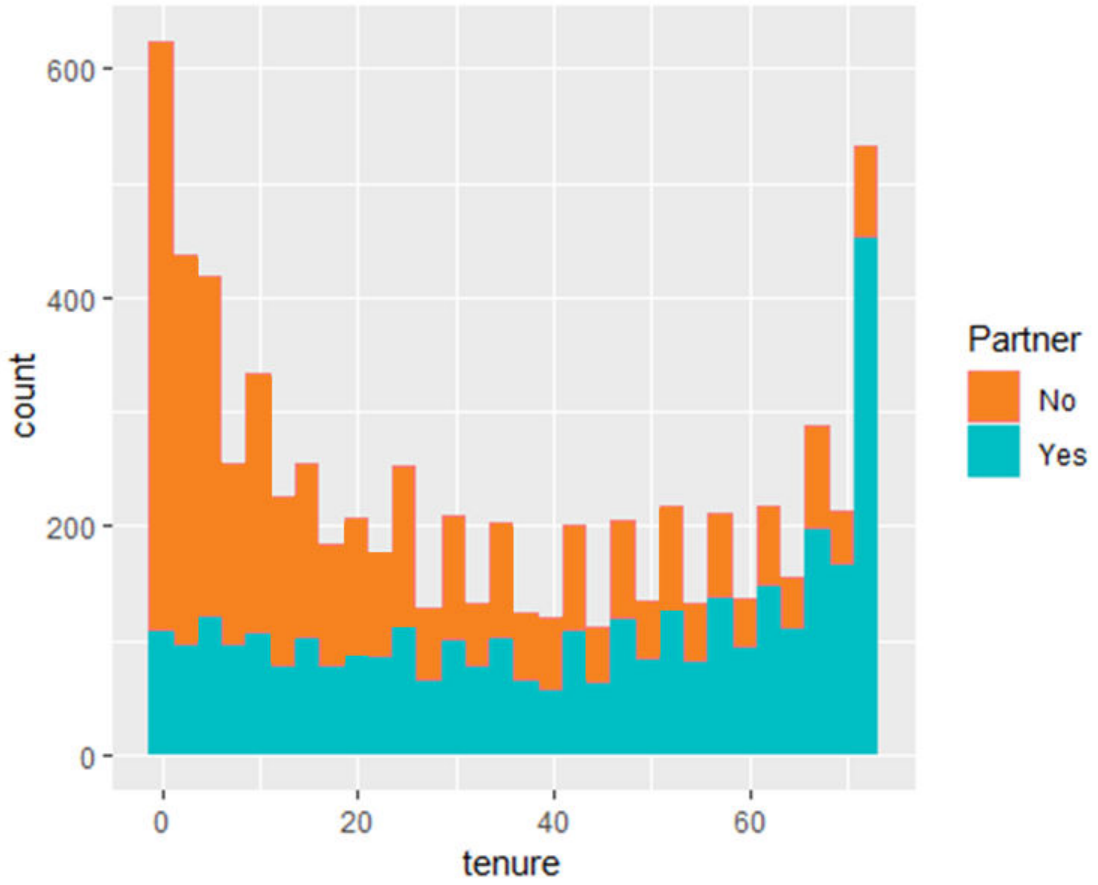
```
ggplot(data = customer_churn, aes(x=tenure))+
geom_histogram(fill="palegreen4", col="green")
```



geom_hist()

Assign ‘Partner’ to fill aesthetic

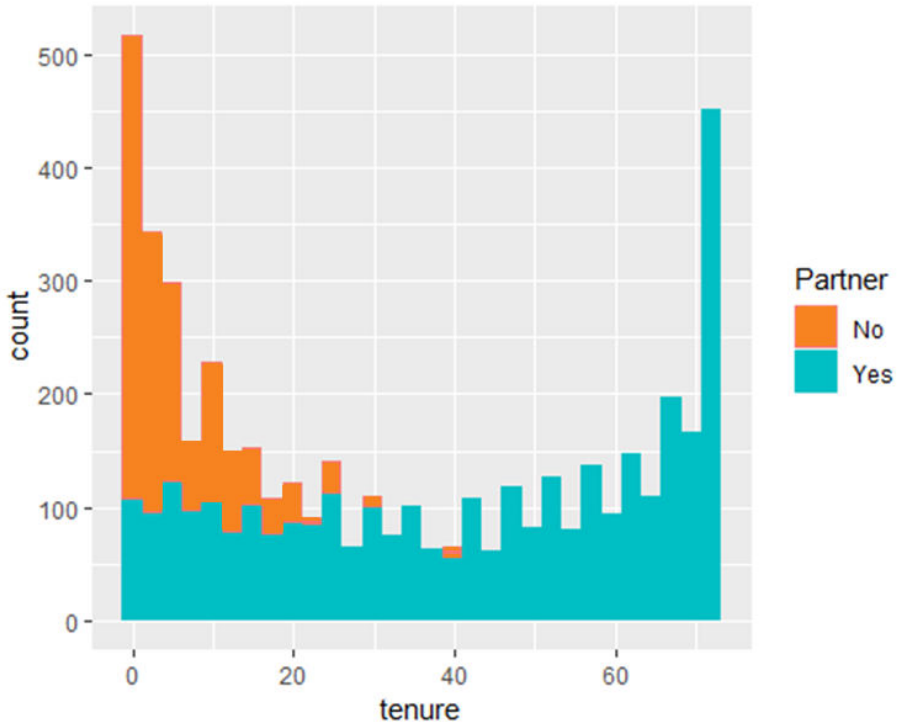
```
ggplot(data = customer_churn, aes(x=tenure,
fill=Partner))+geom_histogram()
```



geom_hist()

Set Position to be ‘identity’

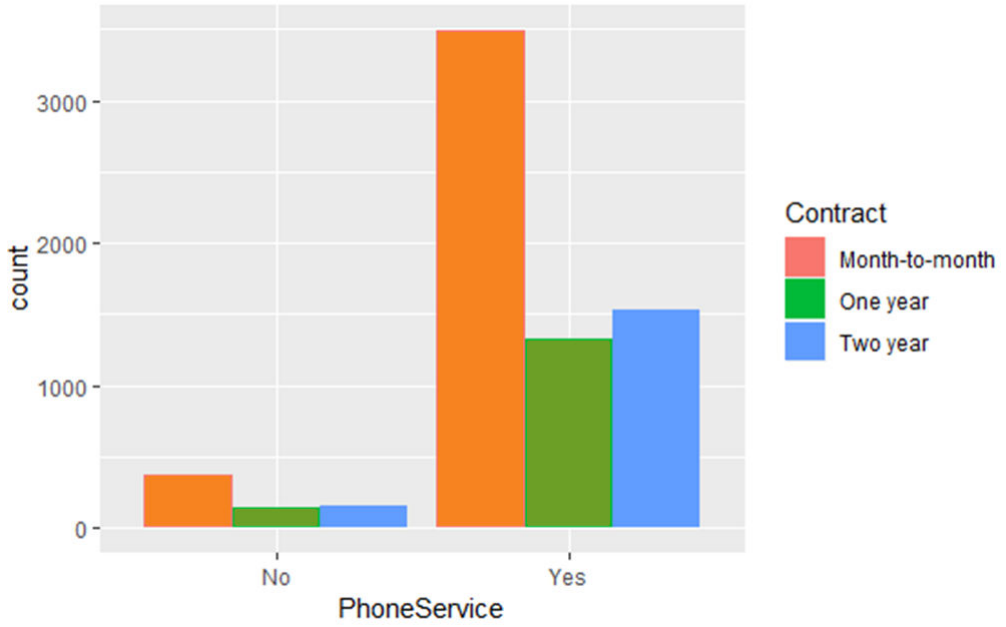
```
ggplot(data = customer_churn,
aes(x=tenure, fill=Partner))+
geom_histogram(position = "identity")
```



geom_bar()

geom_bar()

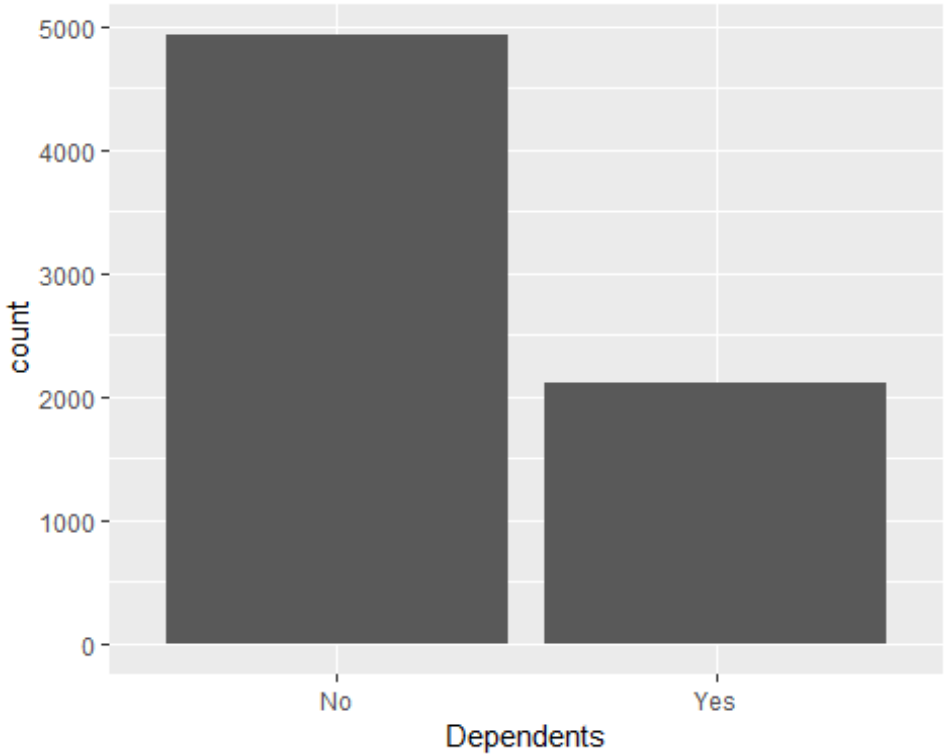
geom_bar() function helps in making bar-plots with ggplot2



geom_bar()

Build a bar-plot for 'Dependents' column

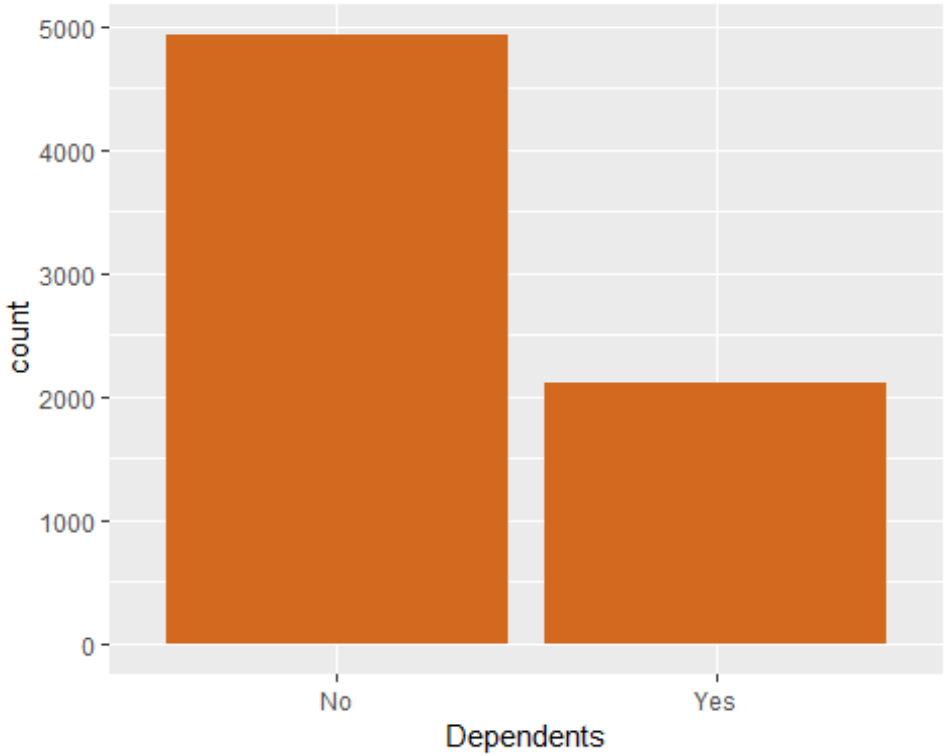
```
ggplot(data = customer_churn,
aes(x=Dependents))+geom_bar()
```



geom_bar()

Add fill color

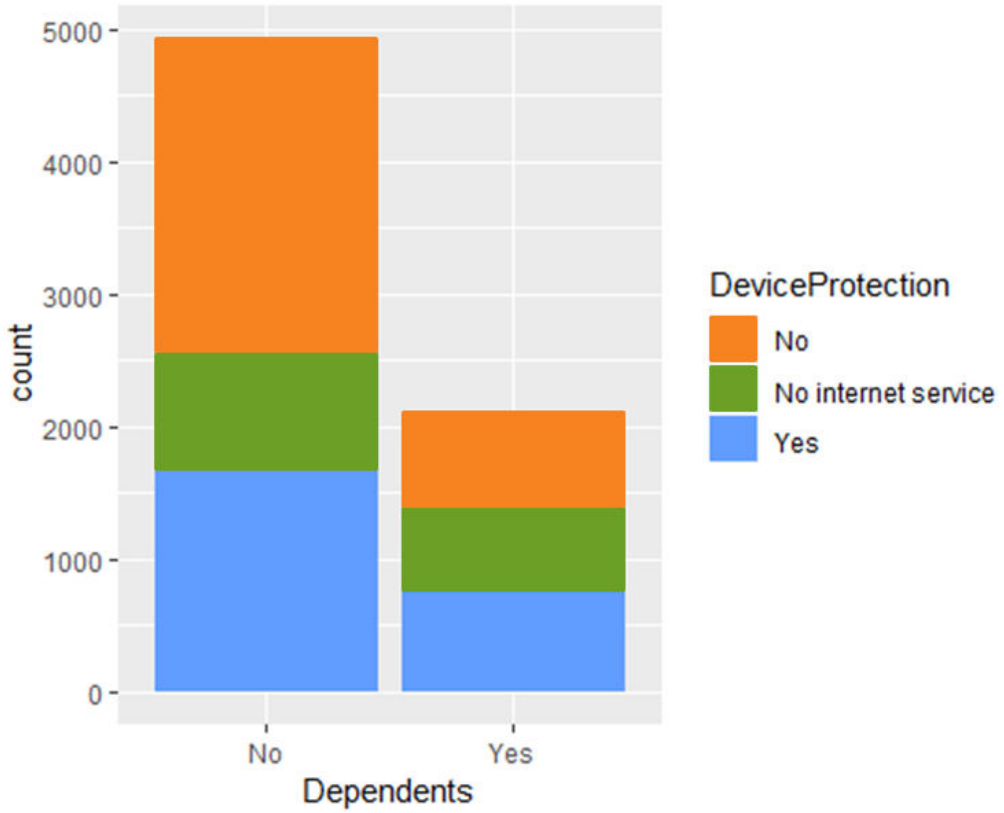
```
ggplot(data = customer_churn,
aes(x=Dependents))+geom_bar(fill="chocolate")
```



geom_bar()

Assigning 'DeviceProtection' column to fill aesthetic

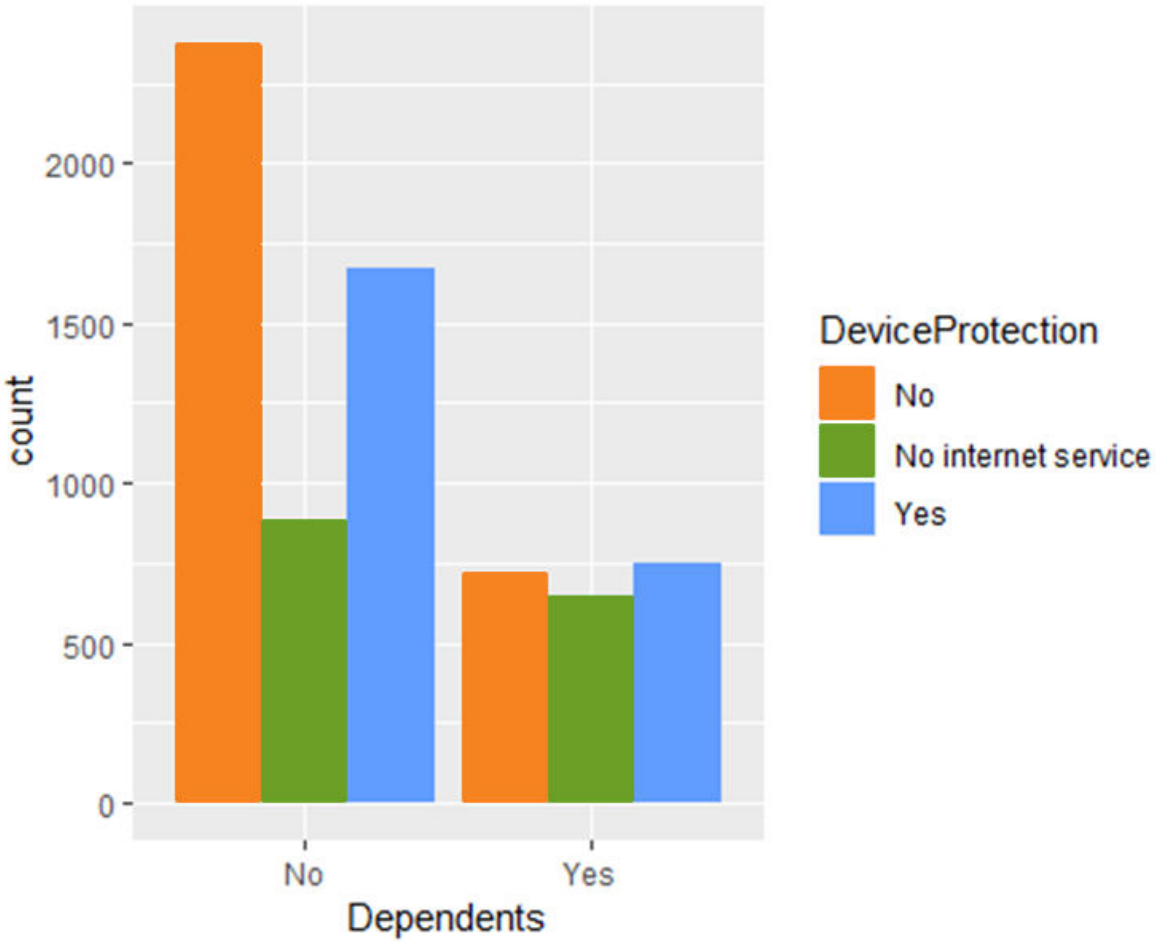
```
ggplot(data = customer_churn,
aes(x=Dependents,fill=DeviceProtection))
+geom_bar()
```



geom_bar()

Set the position to ‘dodge’

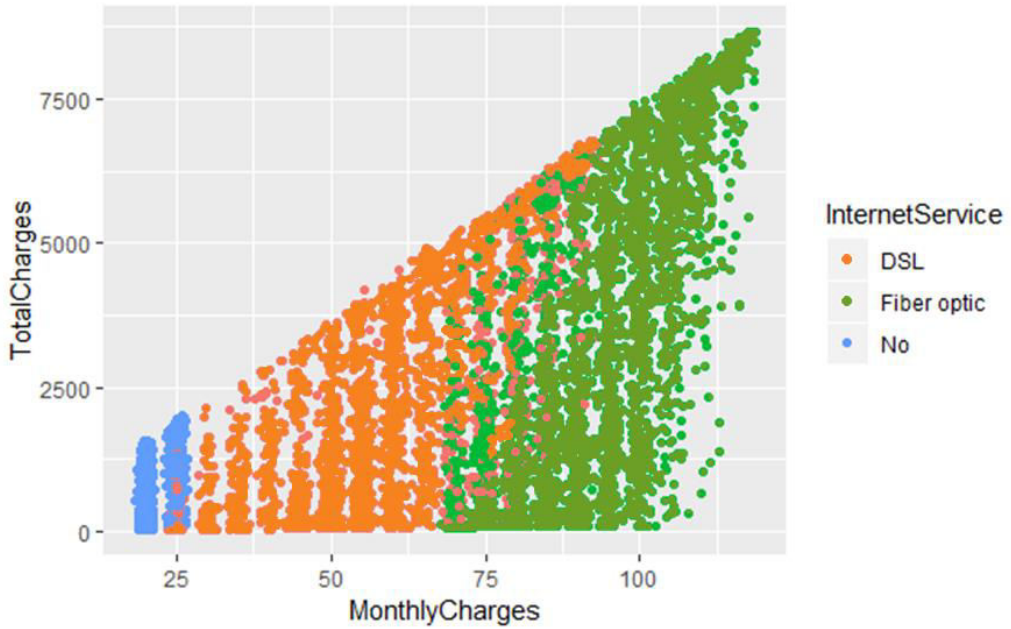
```
ggplot(data = customer_churn,
aes(x=Dependents,fill=DeviceProtection))
+geom_bar(position=‘dodge’)
```



geom_point()

geom_point()

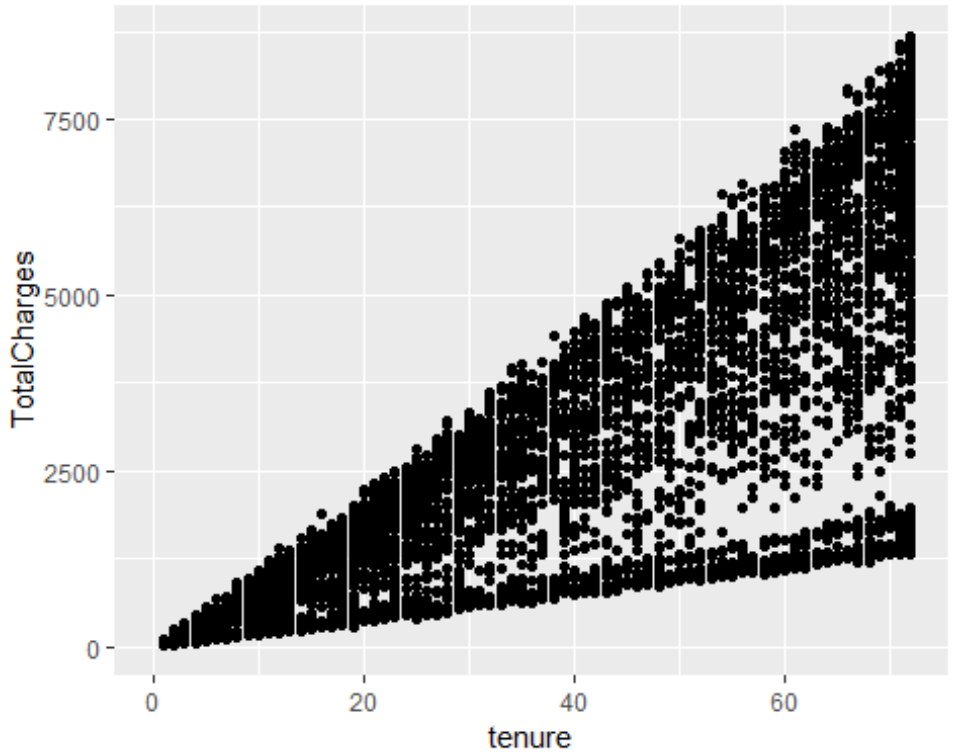
geom_point() function helps in making scatterplots with ggplot2. A scatter plot helps in understanding how does one variable change w.r.t another variable. It is used for two continuous values



geom_point()

Scatter-plot between ‘TotalCharges’ & ‘tenure’

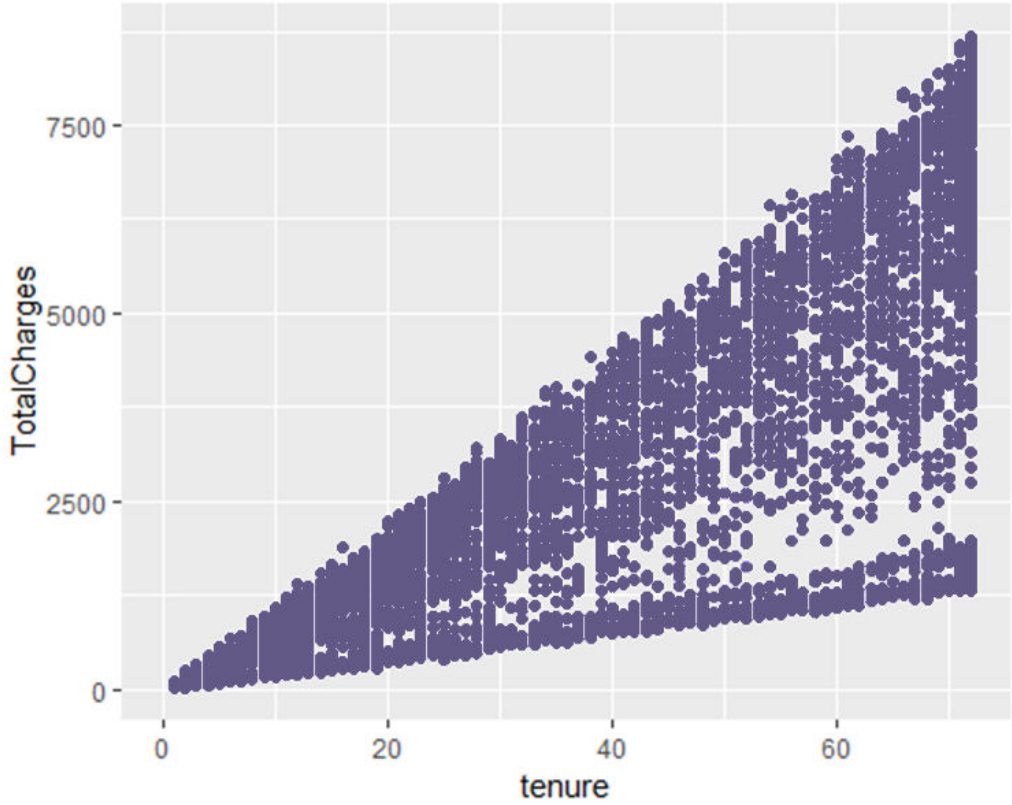
```
ggplot(data = customer_churn,
aes(y=TotalCharges,x=tenure))+geom_point(
)
```



geom_point()

Adding color

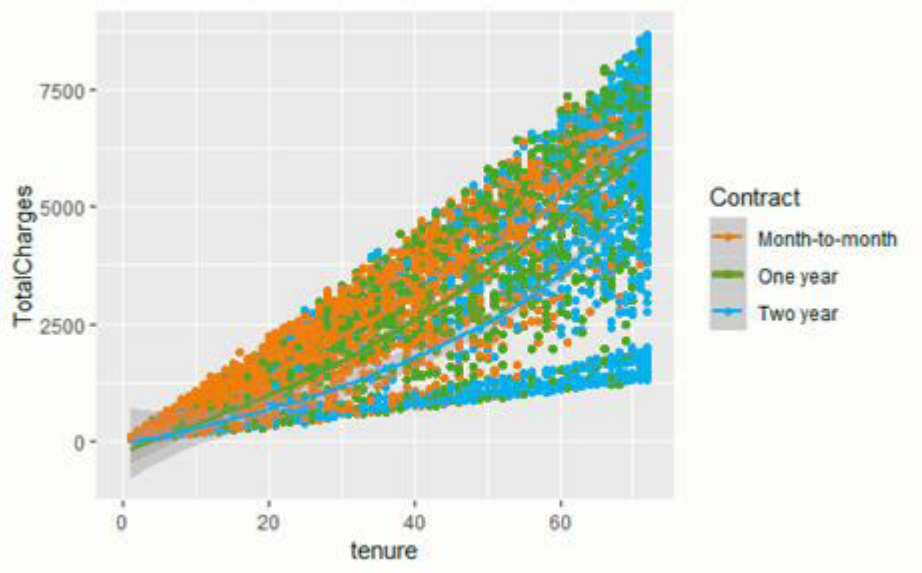
```
ggplot(data = customer_churn,
aes(y=TotalCharges,x=tenure)) +
geom_point(col="slateblue3")
```



geom_point()

Map ‘Partner’ to col aesthetic

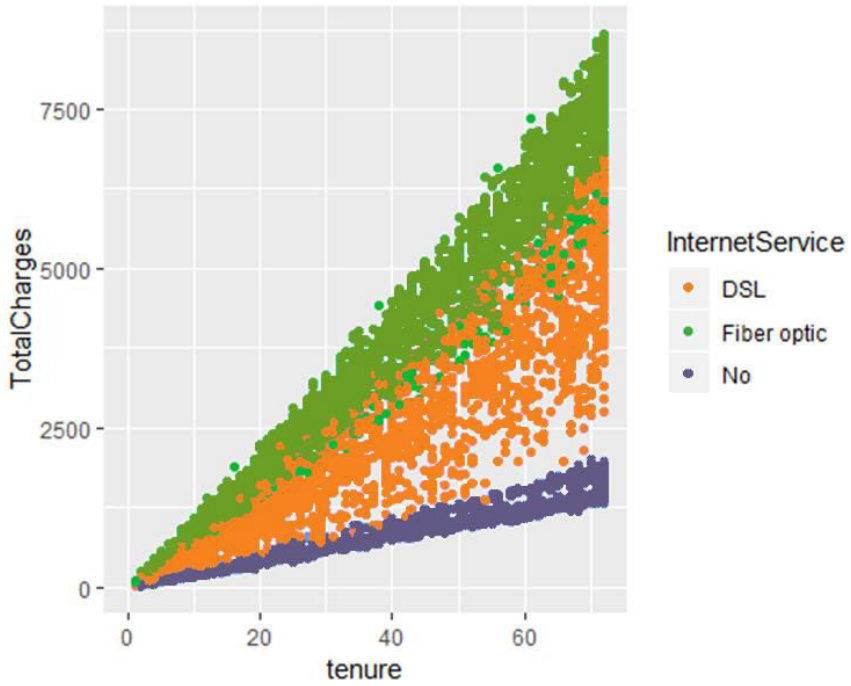
```
ggplot(data = customer_churn,
aes(y=TotalCharges,x=tenure, col=Partner)) +
geom_point()
```



geom_point()

Map 'InternetService' to col aesthetic

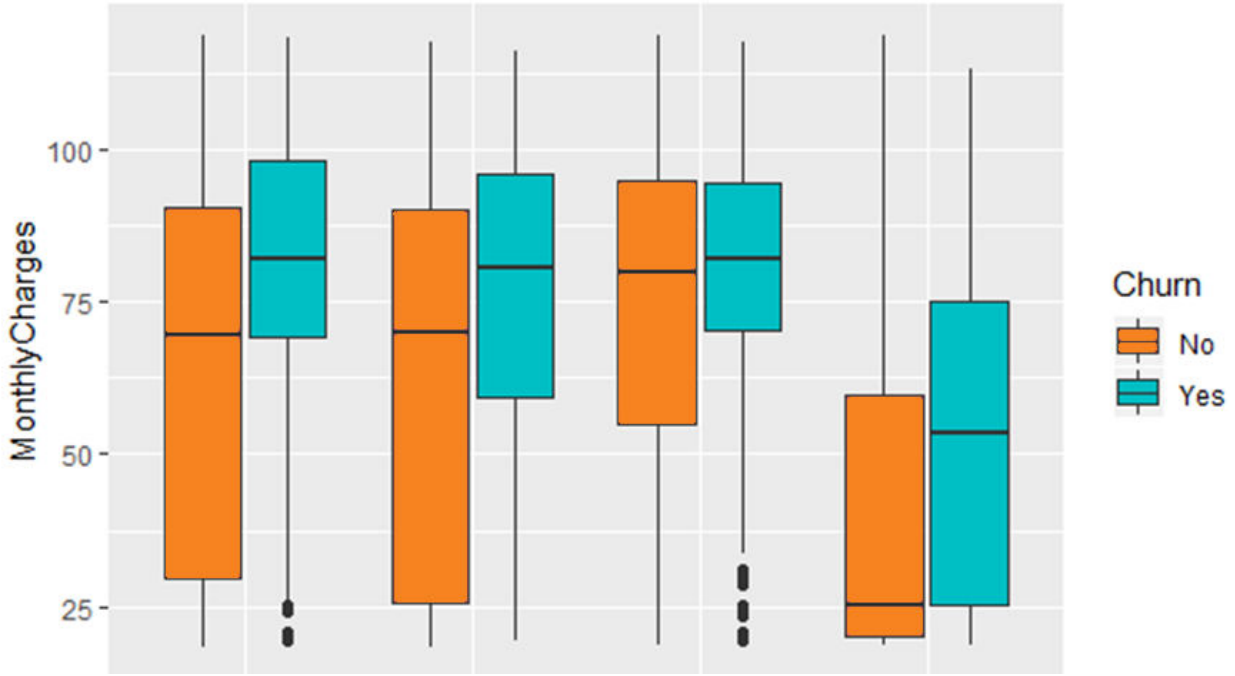
```
ggplot(data = customer_churn,
aes(y=TotalCharges,x=tenure,
col=InternetService)) +
geom_point()
```



geom_boxplot()

geom_boxplot()

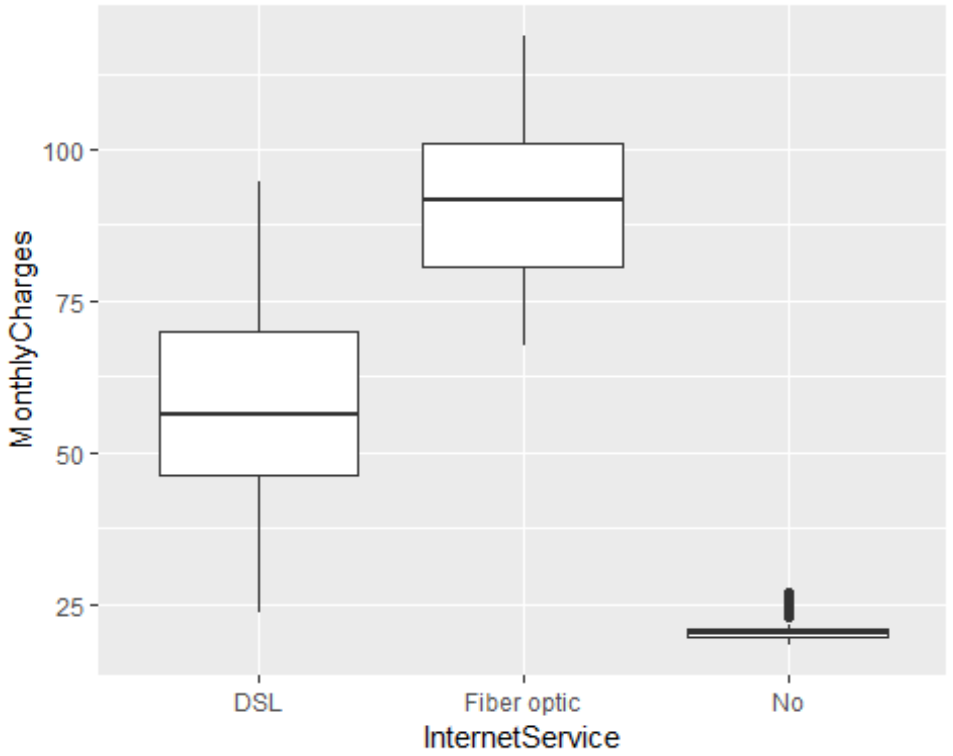
geom_boxplot() function helps in making boxplots with ggplot2. Box Plot shows 5 statistically significant numbers- the minimum, the 25th percentile, the median, the 75th percentile and the maximum. It is thus useful for visualizing the spread of the data and deriving inferences accordingly



geom_boxplot()

Box-plot between ‘MonthlyCharges’ & ‘InternetService’

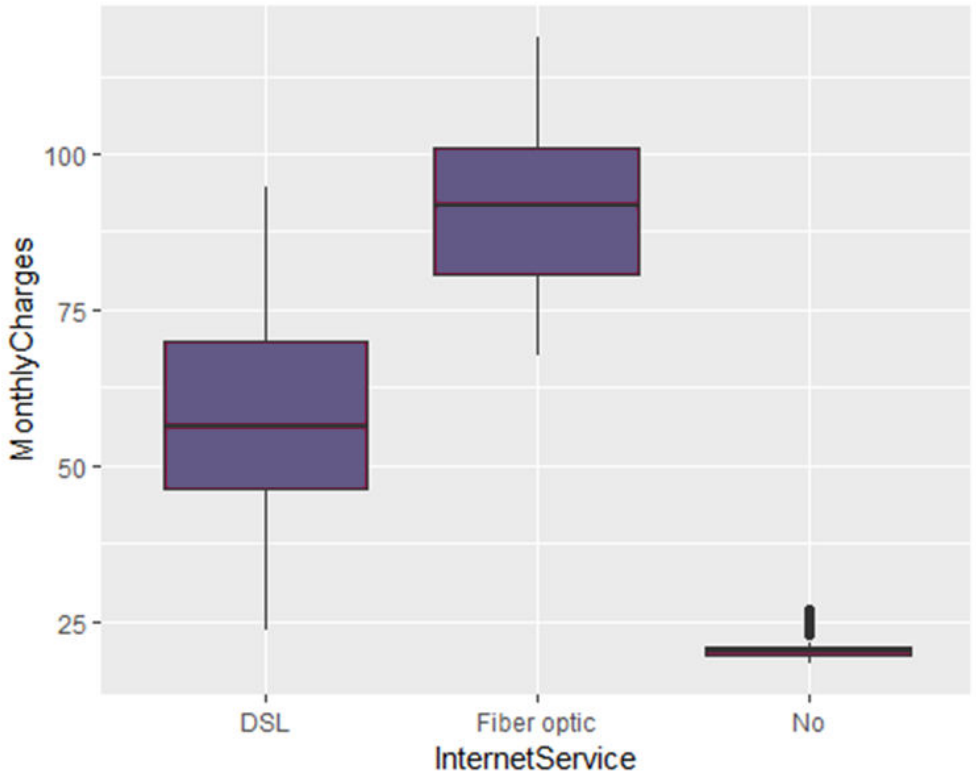
```
ggplot(data = customer_churn,aes(y=MonthlyCharges,x=InternetService))+geom_boxplot()
```



geom_boxplot()

Add fill color

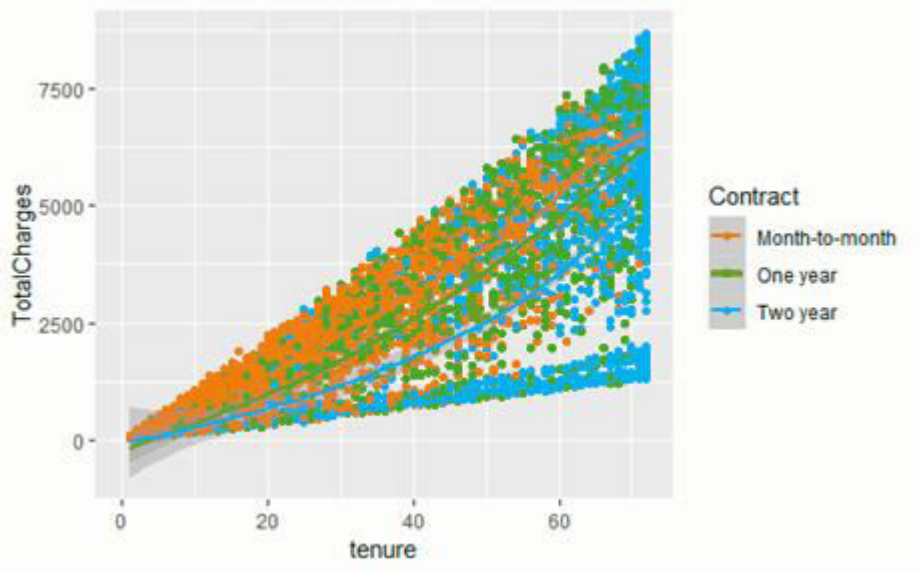
```
ggplot(data =
customer_churn,aes(y=MonthlyCharges,x=Inte
rnetService))+geom_boxplot(fill="violetred4")
```

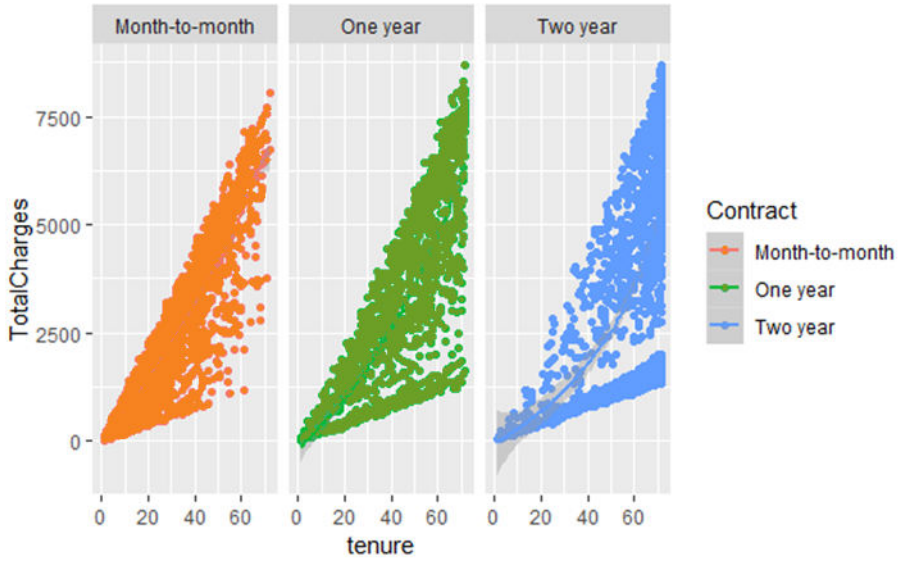


Faceting the data

Faceting the data

facet_grid() is used to facet the data. Faceting is used when the plot is too chaotic and some variables have to be grouped into different facets to have a better visualization

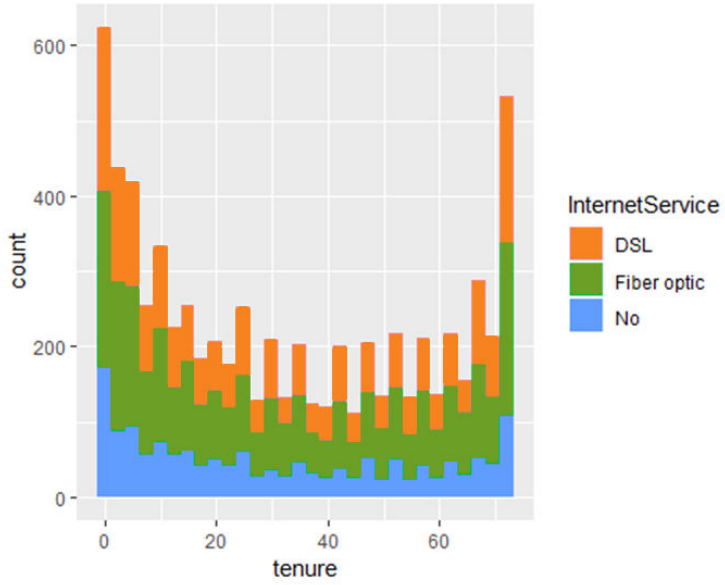




facet_grid()

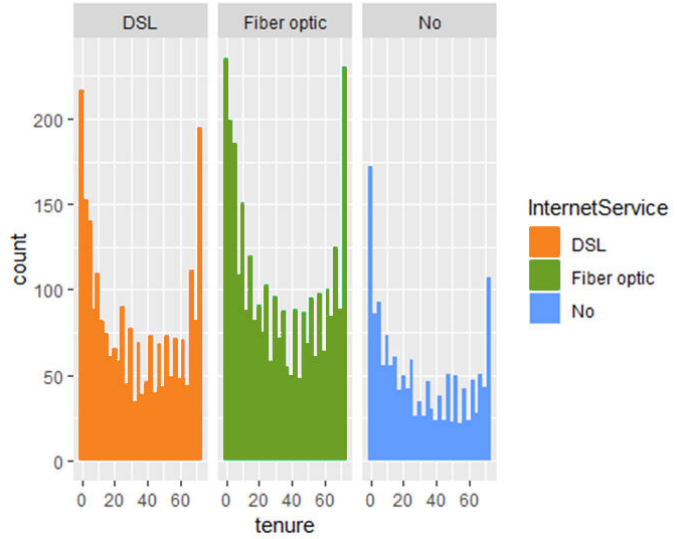
Initial Graph

```
ggplot(data = customer_churn,
aes(x=tenure,fill=InternetService))+
geom_histogram()
```



After Faceting

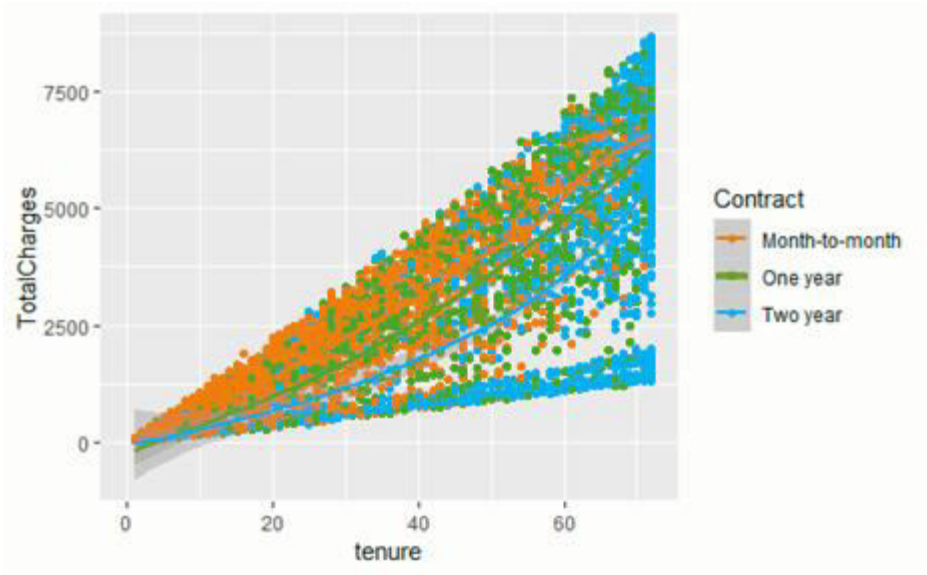
```
ggplot(data = customer_churn,
aes(x=tenure,fill=InternetService))+
geom_histogram()+ facet_grid(~InternetService)
```



facet_grid()

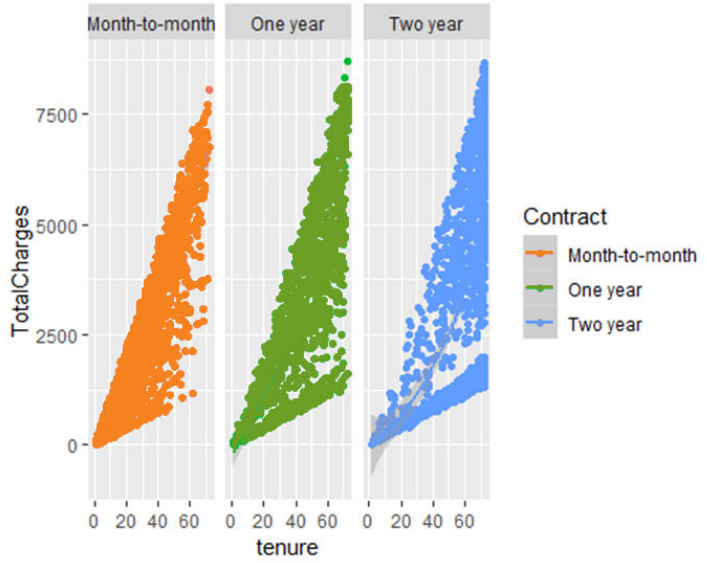
Initial Graph

```
ggplot(data = customer_churn,
aes(y=TotalCharges,x=tenure, col=Contract))+
geom_point()+geom_smooth()
```



After Faceting

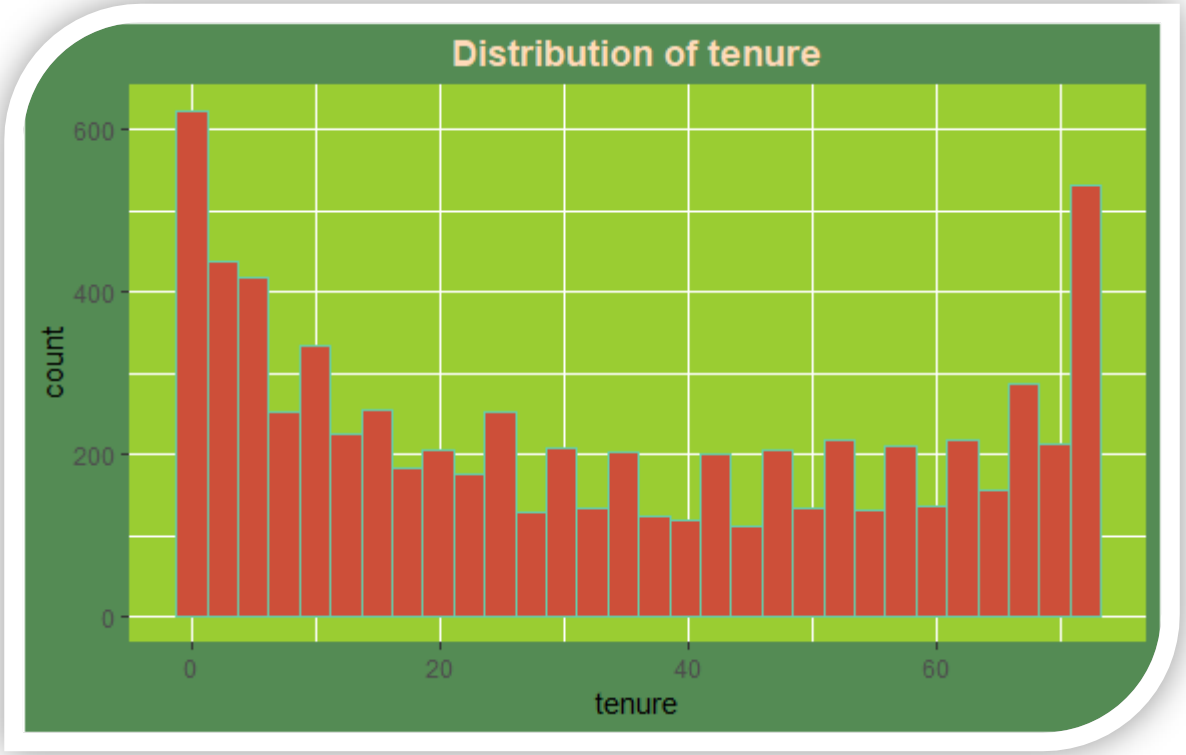
```
ggplot(data = customer_churn,
aes(y=TotalCharges,x=tenure, col=Contract))+
geom_point()+geom_smooth()+facet_grid(~Contract)
```



Theme Layer

Theme Layer

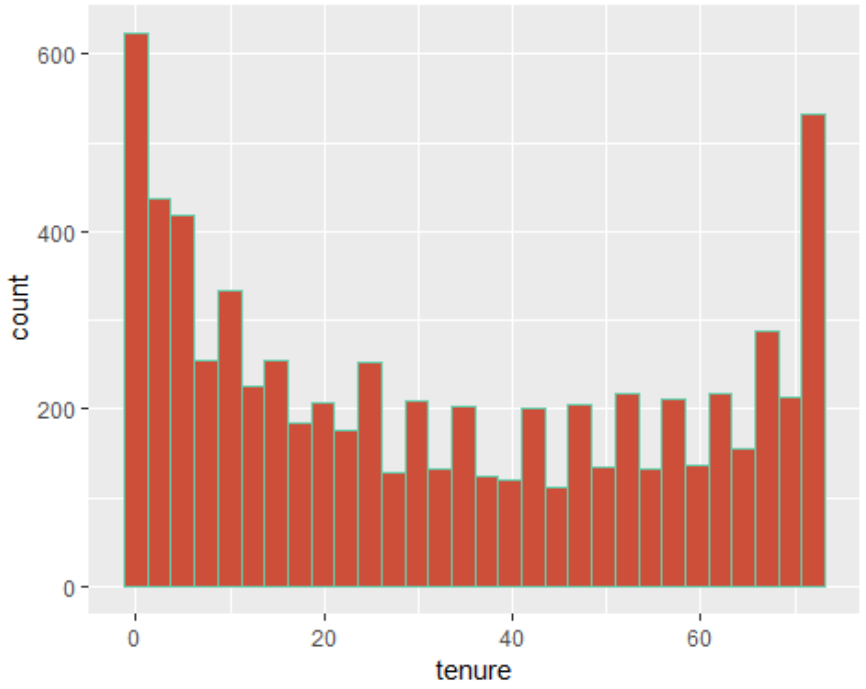
Theme layer is used to add themes to our plots



Theme Layer

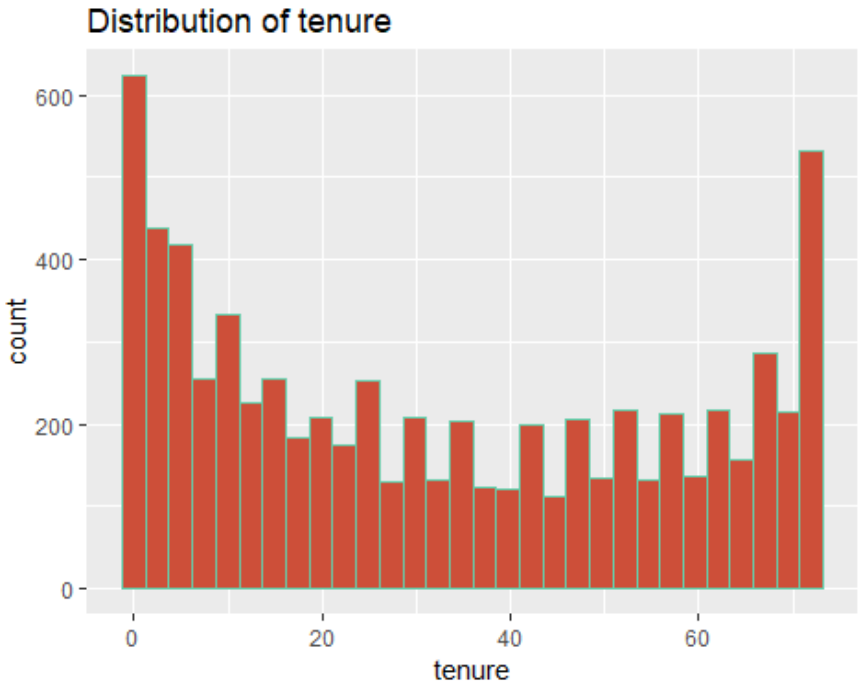
Initial Graph

```
ggplot(data = customer_churn,aes(x=tenure))+
geom_histogram(fill="tomato3",
col="mediumaquamarine") -> g1
```



After Adding Title

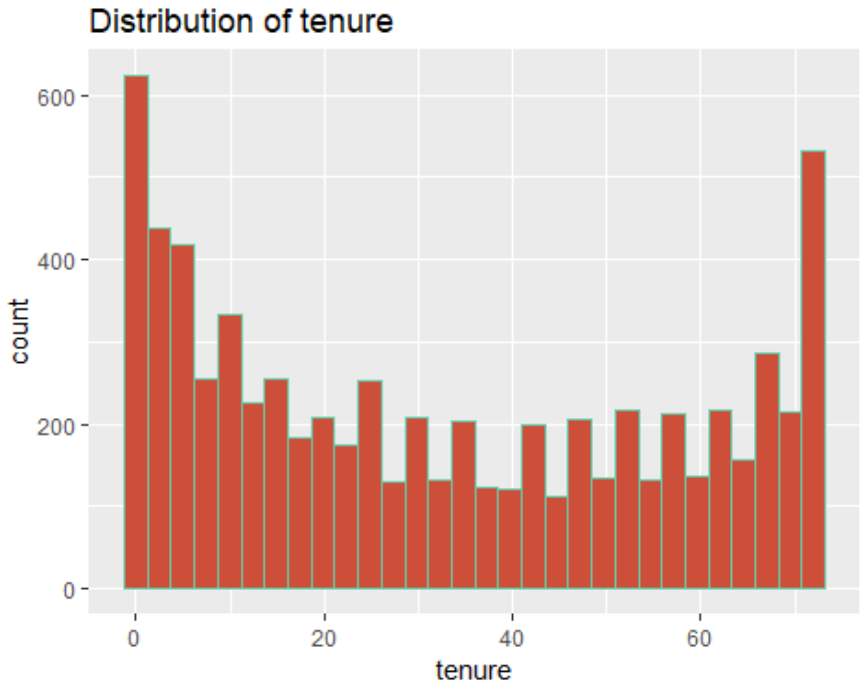
```
g1+labs(title = "Distribution of tenure")->g2
```



Theme Layer

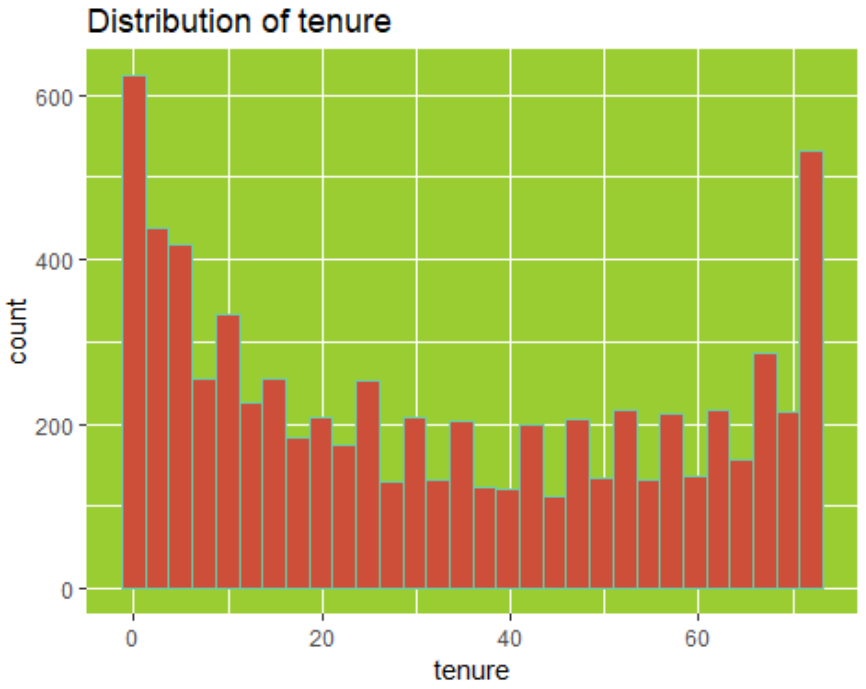
After Adding Title

```
g1+labs(title = "Distribution of tenure")->g2
```



After Adding Panel Background

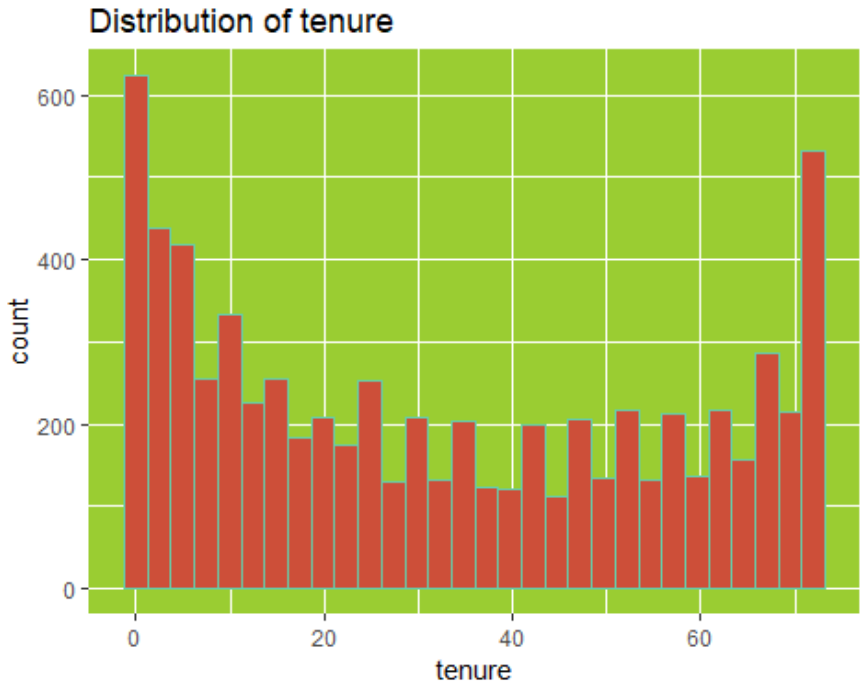
```
g2+theme(panel.background =
element_rect(fill = "olivedrab3"))->g3
```



Theme Layer

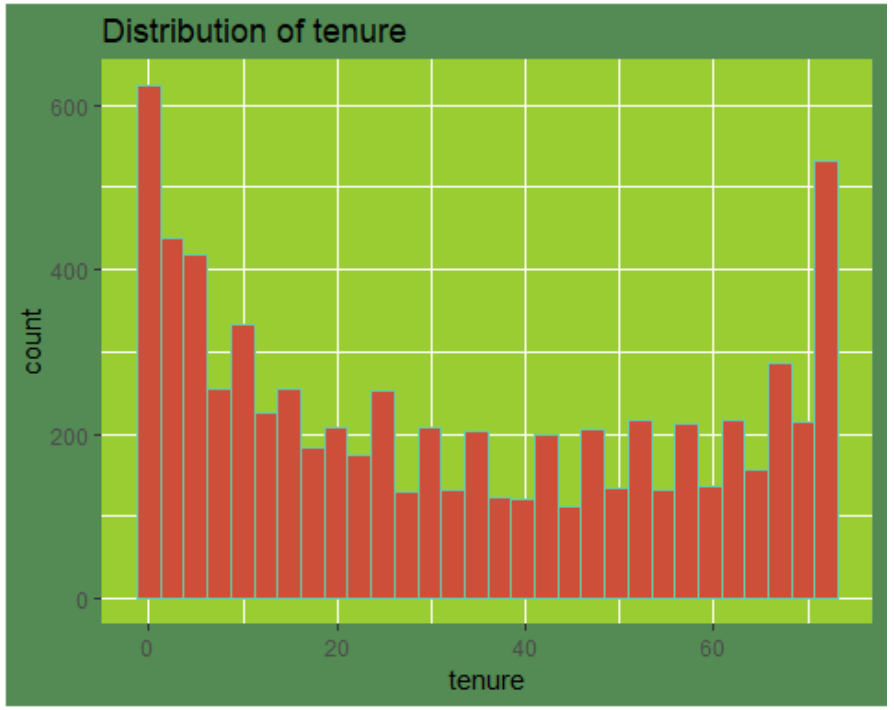
After Adding Panel Background

```
g2+theme(panel.background =
element_rect(fill = "olivedrab3"))->g3
```



After Adding Plot Background

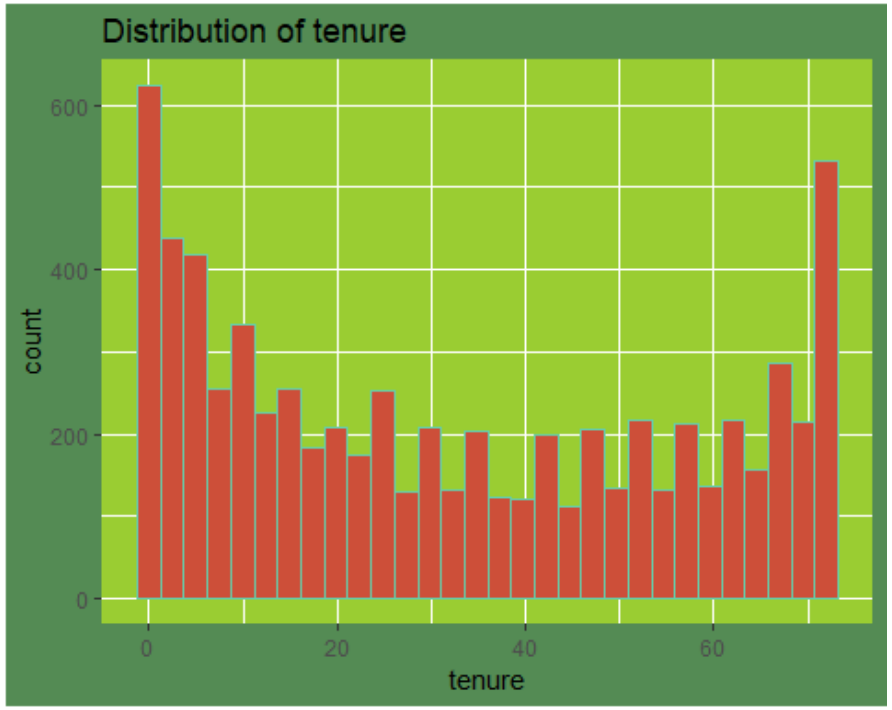
```
g3+theme(plot.background =
element_rect(fill = "palegreen4"))->g4
```



Theme Layer

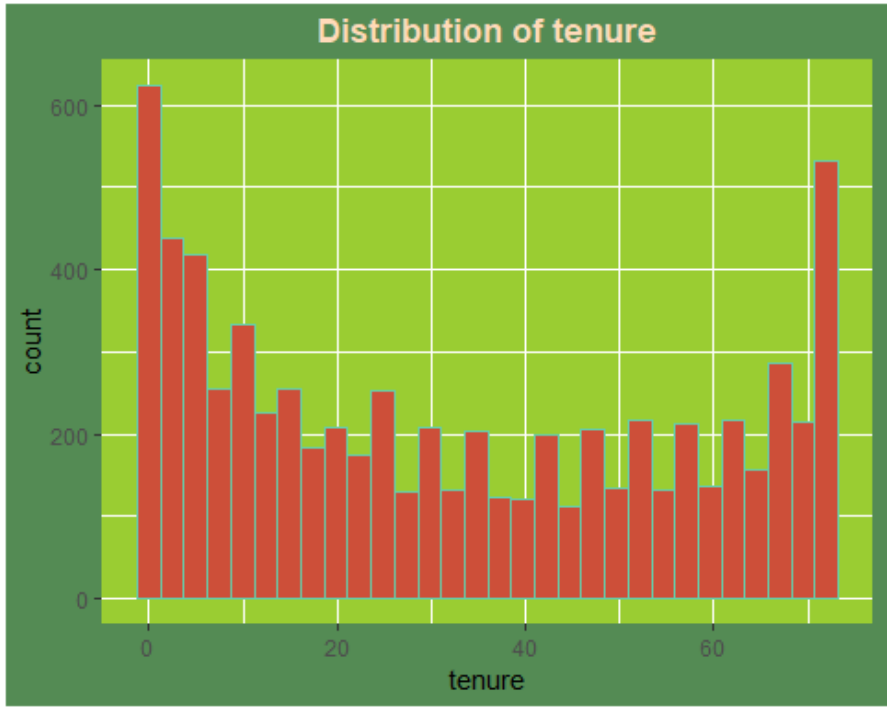
After Adding Plot Background

```
g3+theme(plot.background =
element_rect(fill = "palegreen4"))->g4
```



After Changing Title

```
g4+theme(plot.title = element_text(hjust = 0.5, face="bold",
colour = "peachpuff"))
```



Quiz

Quiz

Which of these is the correct code to make a bar-plot for the 'TechSupport' column. The color of the bars should be 'blue' & the title of the plot should be 'Distribution of Tech Support'

1. plot(customer_churn$TechSupport, col="blue", main="Distribution of Tech Support")
2. plot(customer_churn$TechSupport, fill="blue", main="Distribution of Tech Support")
3. plot(customer_churn$TechSupport, col="blue", title="Distribution of Tech Support")
4. plot(customer_churn$TechSupport, color="blue", title="Distribution of Tech Support")

Quiz

Which of these is the correct code to make a histogram for the ‘tenure’ column. The fill color of the bins should be ‘azure’ & the number of bins should be 87

1. ggplot(data = customer_churn,aes(x=tenure,col='azure'))+geom_histogram(bins=87)
2. ggplot(data = customer_churn,aes(x=tenure))+geom_histogram(col="azure",bins=87)
3. ggplot(data = customer_churn,aes(x=tenure))+geom_histogram(fill="azure",bins=87)
4. ggplot(data = customer_churn,aes(x=tenure,fill='azure'))+geom_histogram(bins=87)

Quiz

Which of these is the correct code to make a bar-plot for the ‘OnlineBackup’ column. The color of the bars should be determined by the ‘PhoneService’ column

1. ggplot(data = customer_churn,aes(fill=OnlineBackup,x=PhoneService))+geom_bar()
2. ggplot(data = customer_churn,aes(y=OnlineBackup,fill=PhoneService))+geom_bar()
3. ggplot(data = customer_churn,aes(x=OnlineBackup))+geom_bar(fill=PhoneService)
4. ggplot(data = customer_churn,aes(x=OnlineBackup,fill=PhoneService))+geom_bar()

Quiz

To which of these geometries can you add the facet_grid()?

1. geom_bar()
2. geom_histogram()
3. geom_point()
4. All of the above

Thank You