


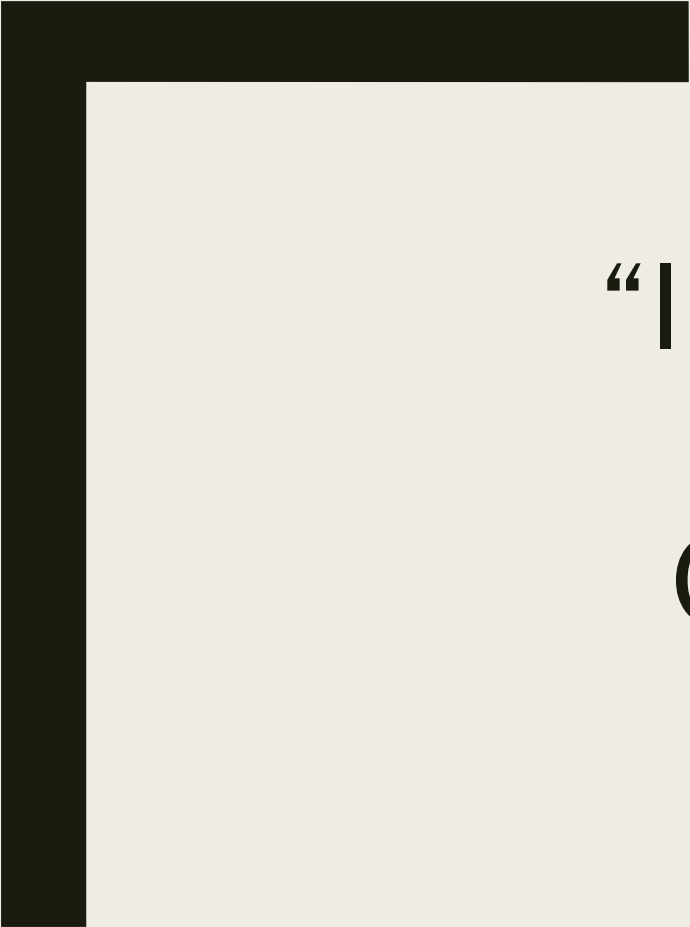


LECTURE 8

BI-VARIATE ANALYSIS

Dr. Jingyuan Zhao





“INTRODUCTION TO STATISTICS” GROUP PROJECT

Dr. Jingyuan Zhao



Stats Project Requirements

- It is a group project

- ☐ Each group with 3-4 students (arrangement by yourself). If your team size is less than 3, there is penalty. If you cannot form a team of size 3, pls email me. I will identify a suitable team for you.
- ☐ Submit your project by 13 Nov (hard deadline, no any further extension)
- ☐ Zip **your slides & R code**, and upload to Luminus\Tools\Files\Stats Project_Submission

- Roles needed in this project

Each role could be handled by multiple students, but all the roles are needed. You could change your role in different sectors. In each sector, I need you sign off with your name and role.

- ☐ **Business Analyst**: understand business knowledge, set up the project scope, hypothesis & objective, tell a story and recommendation
- ☐ **Data Scientist**: design & develop methodology and R code
- ☐ **Code Reviewer**: Review the work that has been done by data scientist

Data Introduction

- The original data set is from the public dataset from UCI Machine learning repository about the residential building. <https://archive.ics.uci.edu/ml/datasets/Residential+Building+Data+Set>
- To align with the content we have learnt from 3 lectures of “Introduction to Statistics”, I have done some changes on the variables. So DON'T surprise if you find the difference.
- In the folder, I provided two sets of data + data description.
 1. *For this project, you focus on “project_residential_price_data.csv “ +”data description.xlsx”. you have 29 variables. The 9th variable “Actual sales price” is the only dependent variables in this study.*
 2. *If you have additional time and interest, you could do one more set of analysis using “project_residential_price_data_optional.csv “ +”data description_optional.xlsx”. I added one more dependent variable (categorical), the 30th variable, high/low margin indicator. You can do analysis for this categorical dependent variable. **However, it is optional!***

Three sectors suggested (but not limited)

Step 1: Descriptive analysis to understand variables

- *Descriptive statistics such as mean, median, variation, correlation, contingency tables.*
- *Visualization using ggplot*

Step 2: Statistical testing to validate your hypothesis

- *Design 2 different types of statistical testing, including all important components*
- *Clearly mention the testing purpose, choose the suitable test with reasoning*
- *Use R conduct statistical testing*
- *Draw conclusion based on your statistical testing results*

Step 3: Predict selling price

- *Experimental design*
- *Prediction and validation – any model is acceptable!*
- *How to improve prediction accuracy?*

Indicative Storyline in PPT

Assume that you will present to business stakeholders. Your analysis is helping them understand the drivers and their impact on residential building sales price, and how to use analysis result to make their decision.

- Propose/assume a business problem statement
- High-level summary of your project including method, deliverables, business benefits and team capabilities, assumptions if any
- Your solutioning with results
- Your recommendations for your business stakeholder

Please remember to make it as a story, not just a technical report!

Looking forward!

AGENDA (L8)

1

Bi-variate analysis 1:
Two categorical variables

2

Bi-variate analysis 2:
Categorical variable (independent variable) and Numerical variable
(dependent variable)

3

Bi-variate analysis 3:
Two numeric variables

Three types of analysis

- Univariate analysis (L7)
 - *the examination of the distribution of cases on only one variable at a time*
- Bivariate analysis (L8)
 - *the examination of two variables simultaneously*
 - *Purpose: determining the empirical relationship between the two variables*
- Multivariate analysis (L9)
 - *the examination of more than two variables simultaneously*
 - *Purpose: determining the empirical relationship among the variables*

Bivariate Analysis

- Bivariate analysis focus on the relationship between two variables
- 3 types bivariable analysis
 - *Two categorical variables:* association between cake flavor chosen and sex?
 - *One categorical (independent), one numerical variable (dependent) :* association between sales volume and promotion status
 - *Two numerical variables:* association between sales volume and product price

How to evaluate the association between two variables

- Two categorical variables
 - Bar-chart with subgroup (*Descriptive*)
 - Contingency Table (*Descriptive*)
 - Chi-square Test: to test if two categorical variables are independent (*Inferential*)
- One categorical variable, one numerical variable
 - Side-by-side histogram, boxplot, density plot (*Descriptive*)
 - T-test to test if two means are same (two groups in categorical variable) (*Inferential*)
 - ANOVA test to test if means of more than two groups are the same (>2 groups in categorical variable) (*Inferential*)
- Two numerical variables
 - Scatter plot (*Descriptive*)
 - Correlation coefficient (*Descriptive*)
 - Linear regression model (*Inferential*)

Two categorical variables

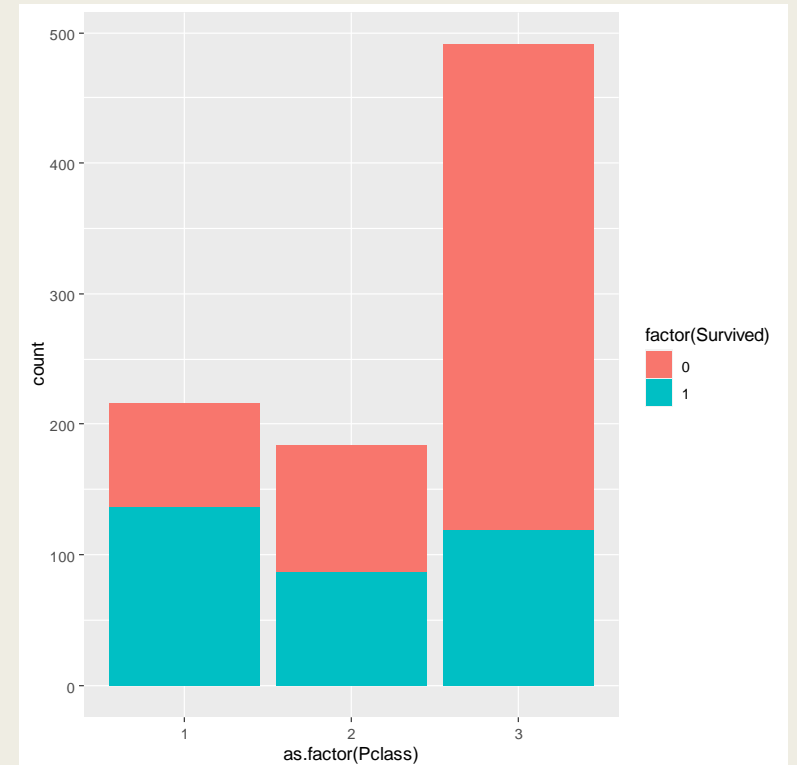
Bar chart with sub-groups for comparison

Let's use Titanic data for R practice of two categorical variables

```
titanic<-read.csv("titanic.csv") #Titanic data
```

```
library(ggplot2)
```

```
ggplot(titanic)+  
aes(x=as.factor(Pclass),fill=factor(Survived))+  
geom_bar()
```



Contingency Tables

- Attributes of independent variable are used as column headings and attributes of the dependent variable are used as row headings

		Medicine Taken		
Cold Length		yes	no	Total
	1 -3 days	86	19	105
	4 - 7 days	16	79	95
	Total	102	98	200

Cold Length associated with Medicine Taken?

Contingency table in R

- `table(..., exclude = if (useNA == "no") c(NA, NaN), useNA = c("no", "ifany", "always"), dnn = list.names(...), deparse.level = 1)`

```
ct<-table(x=titanic$Survived, y=titanic$Pclass) #x: dependent var; y: independent var
```

y

x 1 2 3

0 80 97 372

1 136 87 119

Among all the 1st class passengers, there are 80 survivors.

How to calculate proportion for contingency table in R ?

```
prop.table(x, margin = NULL)
```

X: table

Margin: index, or vector of indices to generate margin for; margin =1 proportion by row, margin = 2 proportion by col

```
prop_ct0<-prop.table(ct)      # by default, overall proportion
```

```
prop_ct1<-prop.table(ct,margin=1) # margin =1, proportion by row
```

```
prop_ct2<-prop.table(ct,margin=2) # margin =2, proportion by col
```

```
> prop_ct0
```

	y				
x	1	2	3		
0	0.08978676	0.10886644	0.41750842		
1	0.15263749	0.09764310	0.13355780		

=1

```
> prop_ct1
```

	y				
x	1	2	3		
0	0.1457195	0.1766849	0.6775956		
1	0.3976608	0.2543860	0.3479532		

=1

```
> prop_ct2
```

	y				
x	1	2	3		
0	0.3703704	0.5271739	0.7576375		
1	0.6296296	0.4728261	0.2423625		

=1

Real-life example: A/B testing in ecommerce



Group A: 1000



Group B: 1000

A/B testing in ecommerce

- Purpose: To understand which design will bring a higher conversion rate.
- The testing results

Customer_ID	Group_ID	Conversion
1	A	0
2	A	1
...
1000	A	1
1001	B	1
...
2000	B	0



You could consider

Group ID & Conversion are two categorical variables

- How to use contingency table to do the initial analysis?

A/B testing in ecommerce

Scenario 1

	Group A	Group B	Total
1	100	200	300
0	900	800	1700
Total	1000	1000	2000

Question: How you tell if Group A and Group B have the same conversion rate?



Conversion_A = 10%
Conversion_B = 20%

Scenario 2

	Group A	Group B	Total
1	100	110	210
0	900	890	1790
Total	1000	1000	2000



Conversion_A = 10%
Conversion_B = 11%

Statistical test

- A **statistical test** is a way to evaluate the evidence the data provides against a hypothesis.
- The intent is to determine whether there is enough evidence to "reject" a null hypothesis.
- Examples of null hypotheses versus alternative challenging hypotheses
 - *H0: The conversion is the same for two designs VS H1: Design B has **a higher conversion than A** (E-commerce) **One-sided test***
 - *H0: the insulin rate of patients receiving a placebo **is equal to** the insulin rate of patients receiving a medication. VS H1: the insulin rate of patients receiving a placebo **is different from** the insulin rate of patients receiving a medication. (Healthcare) **Two-sided test***

Test Statistics

- A test statistic is a random variable that is calculated from sample data and used in a hypothesis test.
- Test statistics to determine whether to reject the null hypothesis.
- The test statistic compares your data with what is expected under the null hypothesis.
- Suppose that T is a given test statistic. t is the value of T given the sample data
- **P-value** the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct

- $\Pr(T \geq t|H)$ for a one-sided (right tail) test,
- $\Pr(T \leq t|H)$ for a one-sided (left tail) test,
- $2 \min\{\Pr(T \leq t|H), \Pr(T \geq t|H)\}$ for a two-sided test,

Chi-square test statistic

The test statistic for the Chi-Square Test of Independence is denoted X^2 , and is computed as:

$$\chi^2 = \sum_{i=1}^R \sum_{j=1}^C \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where

o_{ij} is the observed cell count in the i^{th} row and j^{th} column of the table

e_{ij} is the expected cell count in the i^{th} row and j^{th} column of the table, computed as

$$e_{ij} = \frac{\text{row } i \text{ total} * \text{col } j \text{ total}}{\text{grand total}}$$

The quantity $(o_{ij} - e_{ij})$ is sometimes referred to as the *residual* of cell (i, j) , denoted r_{ij} .

The calculated X^2 value is then compared to the critical value from the X^2 distribution table with degrees of freedom $df = (R - 1)(C - 1)$ and chosen confidence level. If the calculated X^2 value > critical X^2 value, then we reject the null hypothesis.

Scenario 2

	Group A	Group B	Total
1	100	110	210
0	900	890	1790
Total	1000	1000	2000

Test statistic value = 0.5

Scenario 1

	Group A	Group B	Total
1	100	200	300
0	900	800	1700
Total	1000	1000	2000

$O_{11}=100$

$e_{11}= 1000*300/2000=150$

$R=2$

$C=2$

$df = (2-1)*(2-1)=1$

Under the null hypothesis, test statistic follows chi-square distribution with $df=1$

Test statistic value =
 $(100-150)^2/150+(900-850)^2/850+(200-150)^2/150+(800-850)^2/850$
= 39.2

Degree of Freedom	Probability of Exceeding the Critical Value								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38
Not Significant								Significant	

Scenario 1

Test statistic = 39.2

P-value < 0.01

Scenario 2

Test statistic = 0.5

P-value is a little bit less than 0,5

- Significance level:

When setting up a study, a risk threshold above which H_0 should not be rejected must be specified. This threshold is referred to as the **significance level alpha** and should lay between 0 and 1.

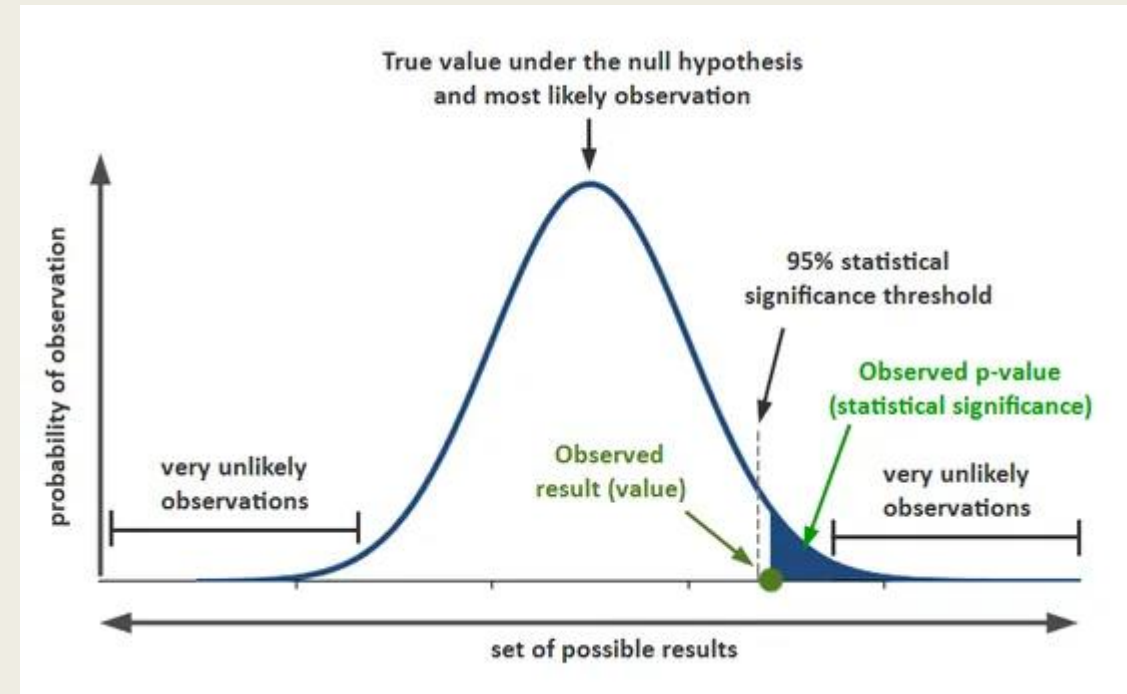
By default, $\alpha = 0.05$

More practically, the p-value should be compared to alpha:

- If **p-value < alpha**, we reject H_0 and accept H_a with a risk proportional to p-value of being wrong.
- If **p-value > alpha**, we do not reject H_0 , but this does not necessarily imply that we should accept it. It either means that H_0 is true, or that H_0 is false but our experiment and statistical test were not “strong” enough to lead to a p-value lower than alpha.

Scenario 1, P-value < 0.01, reject H_0

Scenario 2, P-value > 0.5, do not reject H_0



Bonferroni Correction

- The Bonferroni correction is a multiple-comparison correction used when several dependent or independent statistical tests are being performed simultaneously (since while a given alpha value alpha may be appropriate for each individual comparison, it is not for the set of all comparisons). In order to avoid a lot of spurious positives, the alpha value needs to be lowered to account for the number of comparisons being performed.

Bonferroni Correction: $\frac{\text{significance value}}{\text{number of tests}}$

$$p < 0.05 \rightarrow \text{new } p \text{ value} = \frac{0.05}{5} = 0.01$$

$$p < 0.01 \rightarrow \text{new } p \text{ value} = \frac{0.01}{5} = 0.002$$

$$p < 0.001 \rightarrow \text{new } p \text{ value} = \frac{0.001}{5} = 0.0002$$

Application in Genome-wide association study (GWAS)

consider population structure and family relatedness [3, 4]. Since the publication of MLM for GWAS [3], many MLM-based methods have been developed. All these methods are single-locus, which test one marker at a time, and these methods fail to match the true genetic model of complex traits that are controlled by many loci simultaneously. To overcome this problem, multi-locus models, including FASTmrEMMAa [5], ISIS EM-BLASSO [6], pLARmEB [7], pKWmEB [8], LASSO [9], and FarmCPU [10], have been developed.

Determining the correct P -value threshold for statistical significance is critical to differentiate true positives from false positives and false negatives. To determine the statistical significance threshold in GWAS, different statistical procedures accounting for multiple testing have been proposed, including the Bonferroni correction, Sidak correction, False Discovery Rate (FDR), permutation test, and Bayesian approaches. Bonferroni correction and FDR [11,12,13,14,15] are the two most commonly used methods for crops. All of these methods limit type 1 errors (false-positives), but they almost certainly inflate type 2 errors (false negatives) [16].

Chi-square test

- One of the most common measures of association for contingency tables is Chi-square.
- With this statistic we compute the expected frequencies for the cells which would represent the case that there is no relationship among the variables.
- As the actual numbers depart from the expected values, the larger and more significant Chi-square becomes.
- The significance level of Chi-square depends on the number of observations and the number of cells in the table and so for census data, which often has very large counts, small deviations from the expected values will be statistically significant.
- Chi-square also expects **at least 5 cases** in each cell in order to estimate values reliably.

Chi-square test in R

```
chisq.test(x, y = NULL, correct = TRUE, p = rep(1/length(x), length(x)), rescale.p = FALSE, simulate.p.value = FALSE, B = 2000)
```

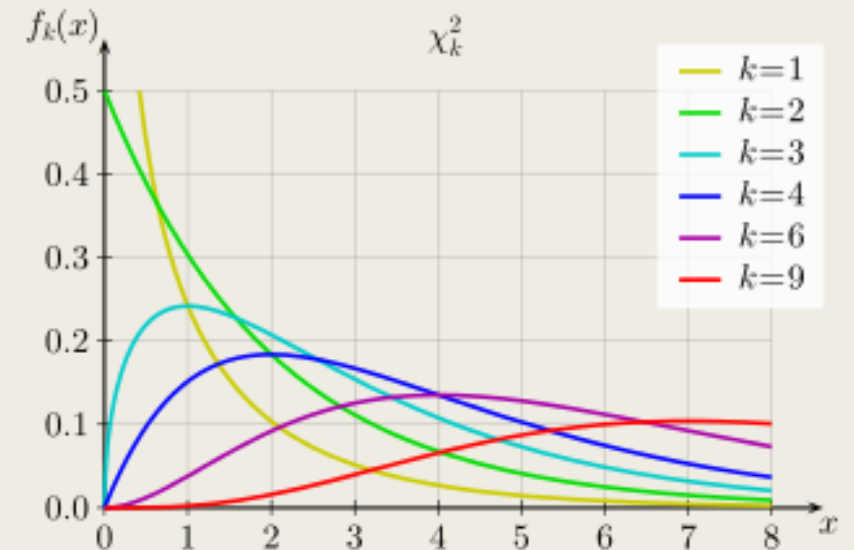
Let's continue R practice with Titanic data

```
chisqt<-chisq.test(x=titanic$Survived, y=titanic$Pclass)
```

Pearson's Chi-squared test

data: titanic\$Survived and titanic\$Pclass

X-squared = 102.89, df = 2, p-value < 2.2e-16



Association between
one categorical with 2 levels (independent var)
and
one numerical (dependent var)

Let's use 'Sales_Jul' data to practice this session.

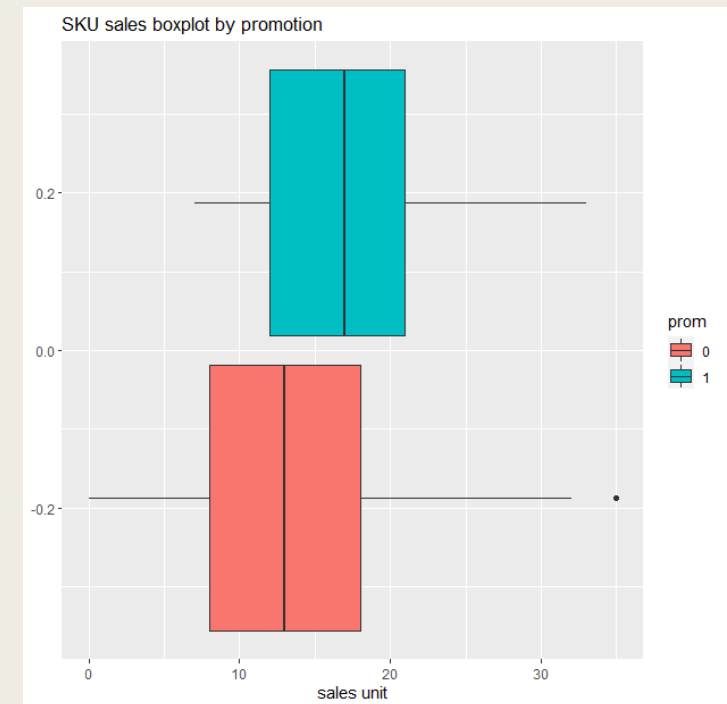
With the purpose of **if sale volume (dependent) is associated with promotion (independent)**

Step 1a. side-by-side density comparison using ggplot2 package

Compare SKU sales by category between with prom and without prom

```
sales_jul$prom<-as.factor(sales_jul$prom) # convert to factor
ggplot(sales_jul)+aes(x=sales_unit,fill=prom)+
  geom_boxplot()+
  labs(x="sales unit",
       title = "SKU sales boxplot by promotion")
```

Step1b. Do self-practice on side-by-side histogram comparison



T-test: Compare the means of two groups

- To determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another.
- H_0 : no true difference between prom and wo prom mean vs. H_1 : have difference (two-sided) OR H_1 : sales mean with promo is higher than without prom (one-sided)
- One sample, two sample, paired?
 - If the groups come from a single population (e.g. measuring before and after an experimental treatment), perform a **paired t-test**.
 - If the groups come from two different populations (e.g. two different species, or people from two separate cities), perform a **two-sample t-test** (a.k.a. **independent t-test**).
 - If there is one group being compared against a standard value (e.g. comparing the acidity of a liquid to a neutral pH of 7), perform a **one-sample t-test**.

T-test statistics

■ Two-sample t-test statistic with the different variance

If the variances of the two groups being compared are different (**heteroscedasticity**), it's possible to use the Welch t-test, which is an adaptation of the Student t-test. The Welch t-statistic is calculated as follow :

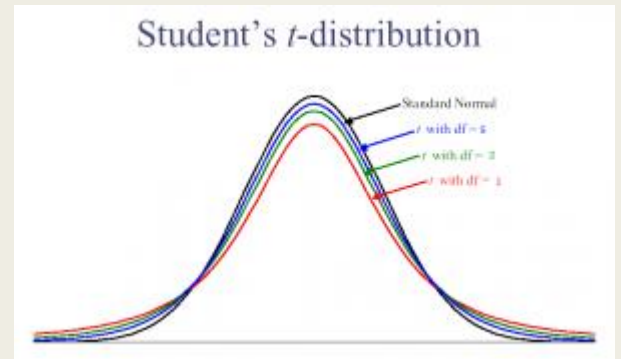
$$t = \frac{m_A - m_B}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}}$$

where, S_A and S_B are the standard deviation of the the two groups A and B, respectively.

Unlike the classic Student's t-test, the Welch t-test formula involves the variance of each of the two groups (S_A^2 and S_B^2) being compared. In other words, it does not use the pooled variance S .

The **degrees of freedom** of **Welch t-test** is estimated as follow :

$$df = \left(\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B} \right)^2 / \left(\frac{S_A^4}{n_A^2(n_A - 1)} + \frac{S_B^4}{n_B^2(n_B - 1)} \right)$$



■ Under the null hypothesis, test statistic follows t-distribution with df as above

T-test in R

```
t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)
```

```
> t.test(sales_unit~prom,data=sales_jul)
```

With difference using two-sided

Welch Two Sample t-test

data: sales_unit by prom

t = -5.9062, df = 204.17, p-value = 1.444e-08

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-4.746107 -2.370424

sample estimates:

mean in group 0 mean in group 1

13.43354 16.99180


```
> t.test(sales_unit~prom,alternative = "less", data=sales_jul)
```

Welch Two Sample t-test

data: sales_unit by prom

t = -5.9062, df = 204.17, p-value = 7.221e-09

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -2.562789

sample estimates:

mean in group 0 mean in group 1

13.43354 16.99180

```
> t.test(sales_unit~prom,alternative = "greater",  
data=sales_jul)
```

Welch Two Sample t-test

data: sales_unit by prom

t = -5.9062, df = 204.17, p-value = 1

alternative hypothesis: true difference in means is
greater than 0

95 percent confidence interval:

-4.553741 Inf

sample estimates:

mean in group 0 mean in group 1

13.43354 16.99180

Association between
one categorical with >2 levels (independent var)
and
one numerical (dependent var)

Independent variable with more than 2 levels

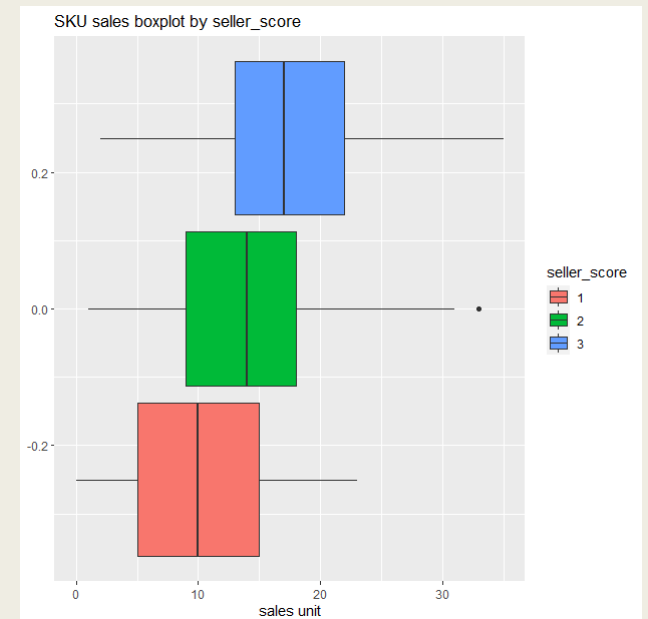
Analyze if seller score cause difference on sales volume

- Seller score: 1,2,3
- ##boxplot using ggplot

```
sales_jul$seller_score<-as.factor(sales_jul$seller_score) # convert to factor
```

```
##boxplot using ggplot
```

```
ggplot(sales_jul)+aes(x=sales_unit,fill=seller_score)+  
  geom_boxplot()+  
  labs(x="sales unit",  
       title = "SKU sales boxplot by seller_score")
```



ANOVA test

- ANOVA (Analysis of Variance) is a statistical test used to analyze the difference between the means of more than two groups.
- The hypotheses of interest in an ANOVA are as follows:
 - $H_0: \mu_1 = \mu_2 = \mu_3 \dots = \mu_k$
 - H_1 : Means are not all equal.

where k = the number of independent comparison groups.

ANOVA test statistics

$$F = \frac{\sum n_j (\bar{X}_j - \bar{X})^2 / (k-1)}{\sum \sum (X - \bar{X}_j)^2 / (N-k)}$$

F-statistic is the ratio of two chi-square statistic

n_j = the sample size in the j^{th} group (e.g., $j=1, 2, 3$, and 4 when there are 4 comparison groups), \bar{X}_j is the sample mean in the j^{th} group, and \bar{X} is the overall mean.

Under the null hypothesis, the test statistics follows the F distribution with (degrees of freedom) $df_1 = k-1$, $df_2 = N-k$, where k is the number of groups, N is the total number of samples

The numerator captures between treatment variability (i.e., differences among the sample means) and the denominator contains an estimate of the variability in the outcome.

The test statistic is a measure that allows us to assess whether the differences among the sample means (numerator) are more than would be expected by chance if the null hypothesis is true.

ANOVA test in R

■ `aov(data, formula, ...)`

```
> aov_seller_score<-aov(sales_unit~seller_score,data=sales_jul)
```

```
>
```

```
> summary(aov_seller_score)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
seller_score	2	3084	1542.2	40.83	<2e-16 ***
Residuals	608	22967	37.8		

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Pair-wise ANOVA

- Compute Tukey Honest Significant Differences

Create a set of confidence intervals on the differences between the means of the levels of a factor with the specified family-wise probability of coverage. The intervals are based on the Studentized range statistic, Tukey's 'Honest Significant Difference' method.

- `TukeyHSD(x, which, ordered = FALSE, conf.level = 0.95, ...)`

`TukeyHSD(aov_seller_score)`

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: `aov(formula = sales_unit ~ seller_score, data = sales_jul)`

`$seller_score`

diff lwr upr p adj

2-1 3.922328 2.298514 5.546142 1e-07

3-1 7.338647 5.427384 9.249909 0e+00

3-2 3.416319 1.960501 4.872136 2e-07

Tukey or Bonferroni or other correction

For Pairwise, Tukey is better!

Comparing the Tukey Procedure with the Bonferroni Procedure

The Bonferroni procedure is a good all around tool, but for all pairwise comparisons the Tukey studentized range procedure is slightly better as we show here.

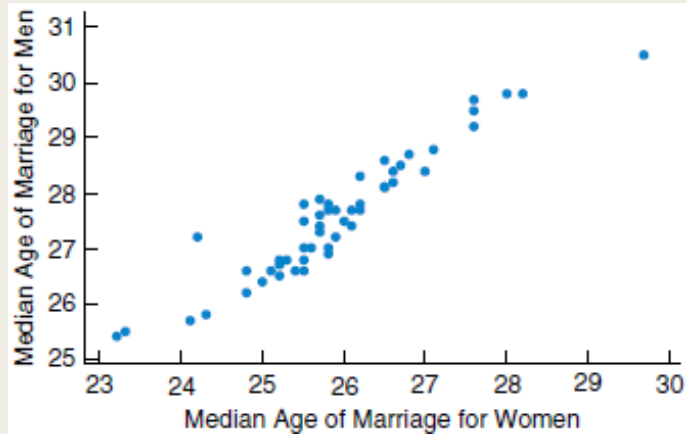
The studentized range is the distribution of the difference between the maximum and a minimum over the standard error of the mean. When we calculate a t -test, or when we're using the Bonferroni adjustment where g is the number of comparisons, we are not comparing apples and oranges. In one case (Tukey) the statistic has a denominator with the standard error of a single mean and in the other case (t -test) with the standard error of the difference between means as seen in the equation for t and q above.

Two Numeric variables

How to analyze the association between 2 numerical variable?

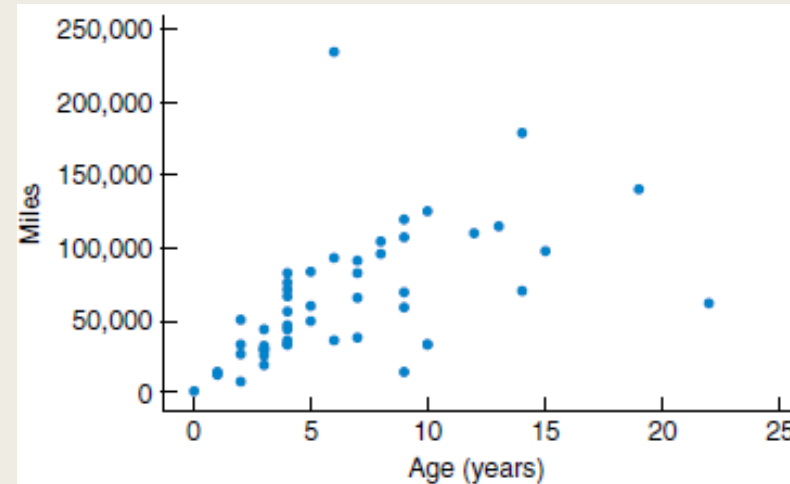
- Descriptive analysis:
 - *Scatter plot*
 - *Correlation*
- Inferential analysis:
 - *Linear Regression*

Scatterplots



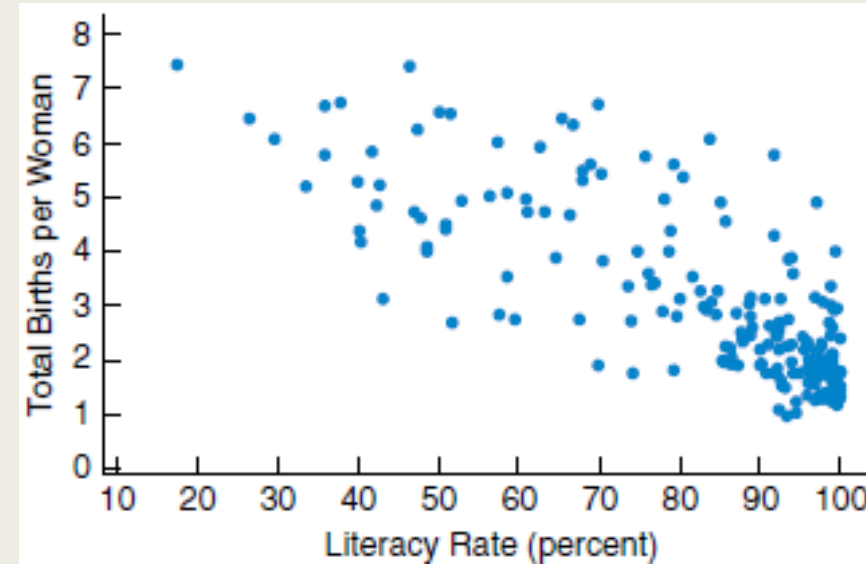
- Used to investigate a positive, negative, or no association between two numerical variables.
- In states where women tend to marry at an older age, men also tend to marry at an older age.

Positive Trend



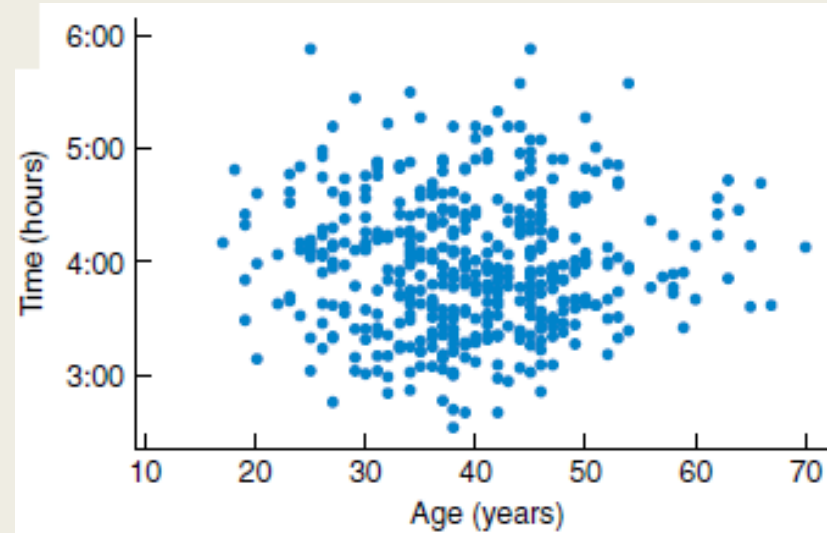
- Older cars tend to have more miles than newer cars.
- Newer cars tend to have fewer miles than older cars.
- There is a **positive association** between car age and miles the car has been driven.

Negative Trend



- Countries with higher literacy rates tend to have fewer births per woman.
- Countries with lower literacy rates tend to have more births per woman.
- There is a **negative association** between literacy rate and births per woman.

No Trend

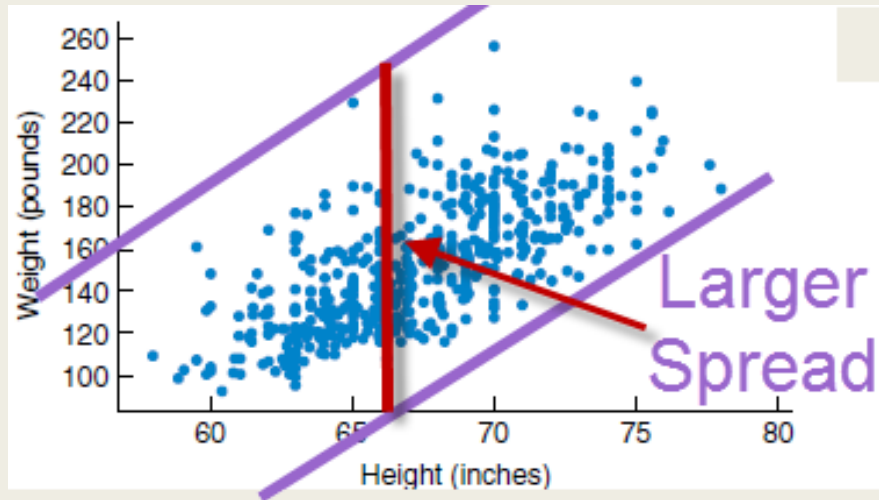


- There is no trend between the speed and age of a marathon runner.
- Knowing the age of a marathon runner does not help predict the runner's speed.
- There is **no association** between a marathon runner's age and speed.

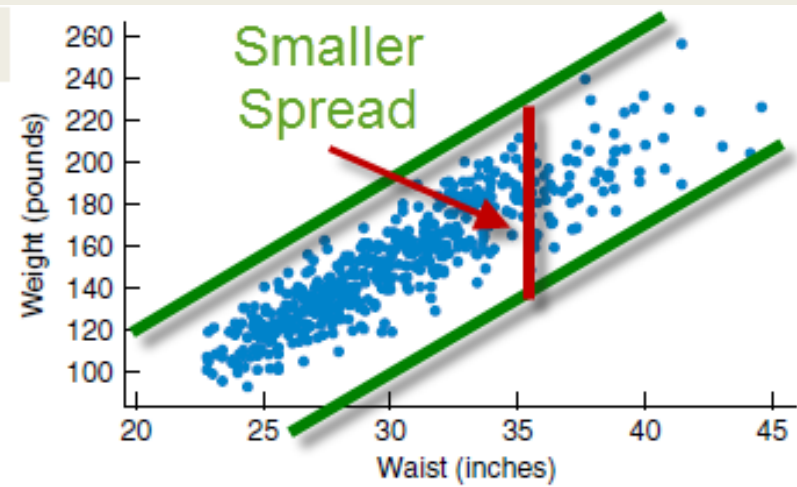
Strength of Association

- If for each value of x , there is a small spread of y values, then there is a **strong association** between x and y .
- If for each value of x , there is a large spread of y values, then there is a **weak or no association** between x and y .
- If there is a **strong** (**weak**) association between x and y , then x is a **good** (**bad**) predictor of y .

Strength of Association

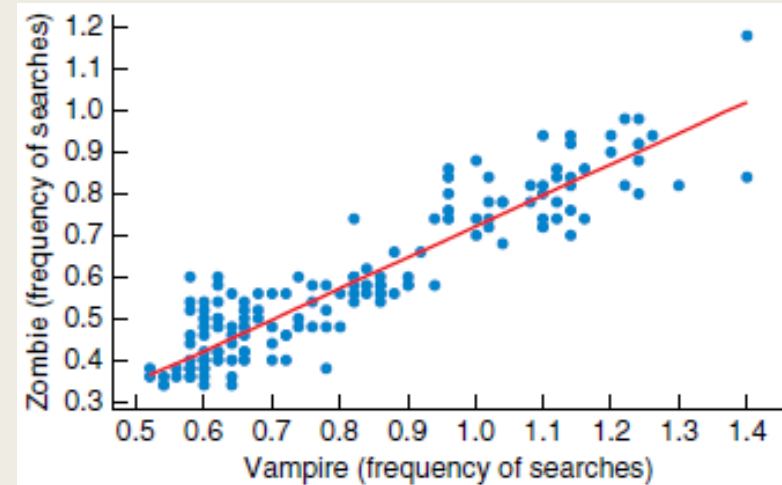


weak association



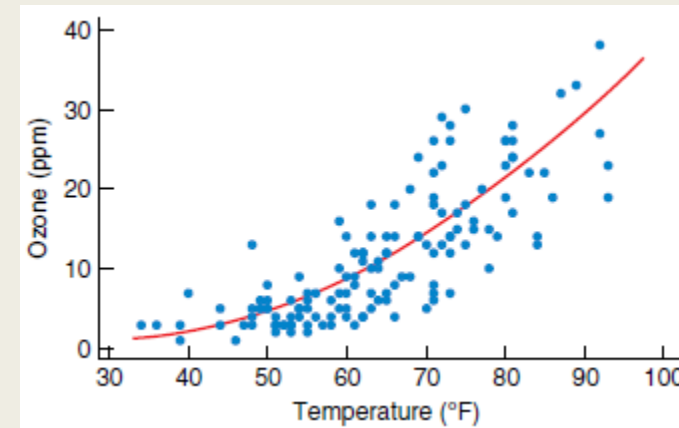
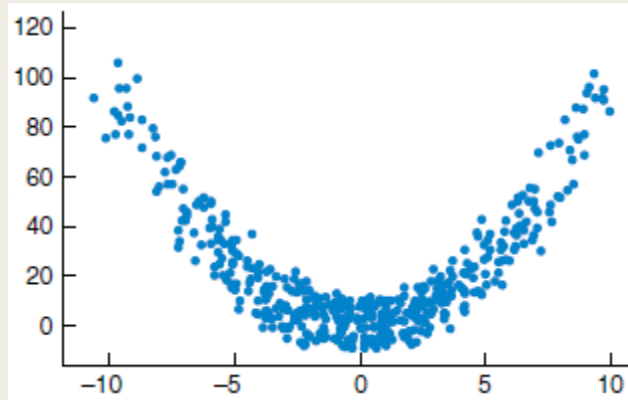
strong association

Linear Trends



- A trend is **linear** if there is a line such that the points in general do not stray far from the line.
- Linear trends are the easiest to work with.
- There is a **positive linear association** between number of searches for “Vampire” and number for “Zombie”.

Other Shapes



- Nonlinear association can also occur, but this is covered in a more advanced statistics course.
- Only use techniques from this chapter when there is a linear trend.

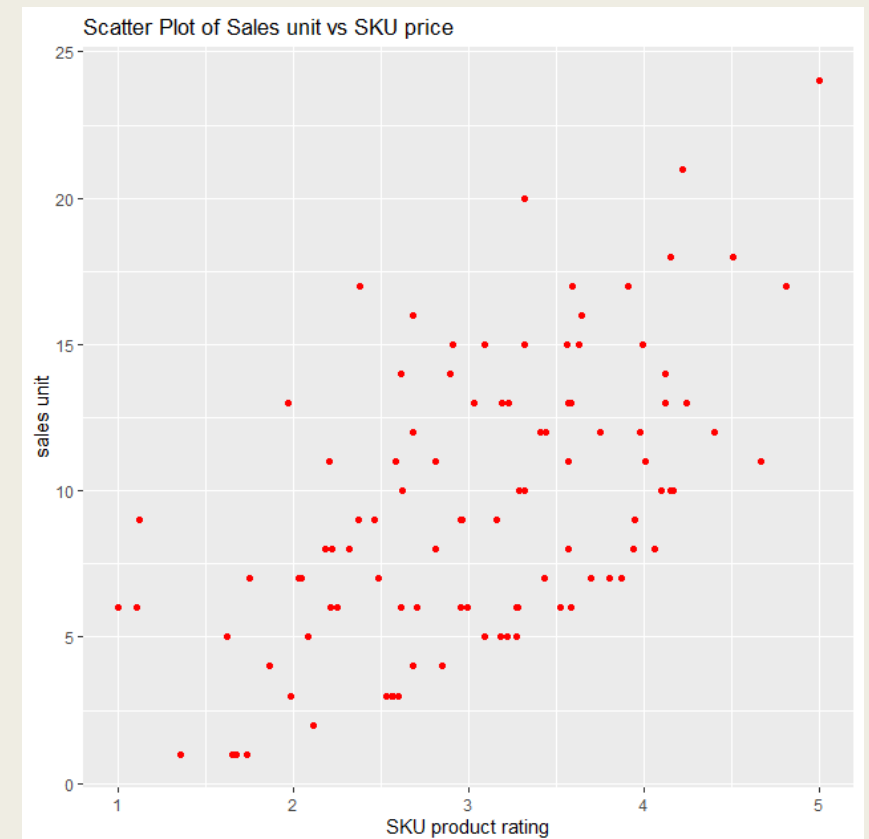
Summary of Analysis of the Scatterplot

- Look to see if there is a trend or association.
- Determine the strength of trend. Is the association strong or weak?
- Look at the shape of the trend. Is it linear? Is it nonlinear?

Scatter plot in R-ggplot2

Variables are mapped to aesthetics with the `aes()` function, while fixed aesthetics are set outside the `aes()` call.

```
ggplot(sales_jul[1:100,])+  
  aes(x=prod_rating,y=sales_unit)+  
  geom_point(col="red")+  
  labs(y="sales unit",  
       x="SKU product rating",  
       title = "Scatter Plot of Sales unit vs SKU price")
```



Scatter plot: set point size

```
ggplot(sales_jul[1:100,])+  
  aes(x=prod_rating,y=sales_unit)+  
  geom_point(col="red", size = 3)+  
  labs(y="sales unit",  
       x="SKU product rating",  
       title = "Scatter Plot of Sales unit vs SKU price")
```



Scatter size: set point shape

```
ggplot(sales_jul[1:100,])+  
  aes(x=prod_rating,y=sales_unit)+  
  geom_point(col="red", size = 3, shape = 17)+  
  labs(y="sales unit",  
       x="SKU product rating",  
       title = "Scatter Plot of Sales unit vs SKU price")
```



Scatter plot with mapping color (1)

- If the variable that maps to color is a factor, then the color scale will change

```
ggplot(sales_jul[1:100,])+
```

```
  aes(x=prod_rating,y=sales_unit,col=as.factor(prom))+
```

```
  geom_point()+
```

```
  labs(y="sales unit",
```

```
        x="SKU product rating",
```

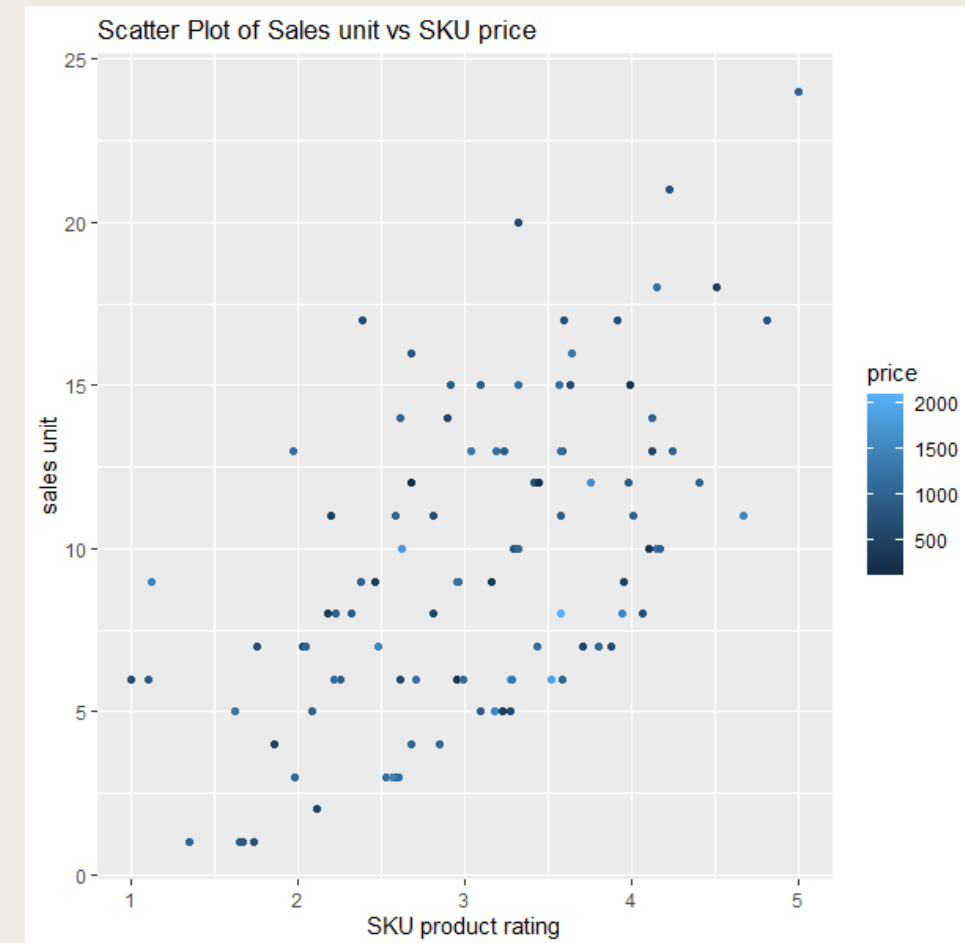
```
        title = "Scatter Plot of Sales unit vs SKU price")
```



Scatter plot with mapping color (2)

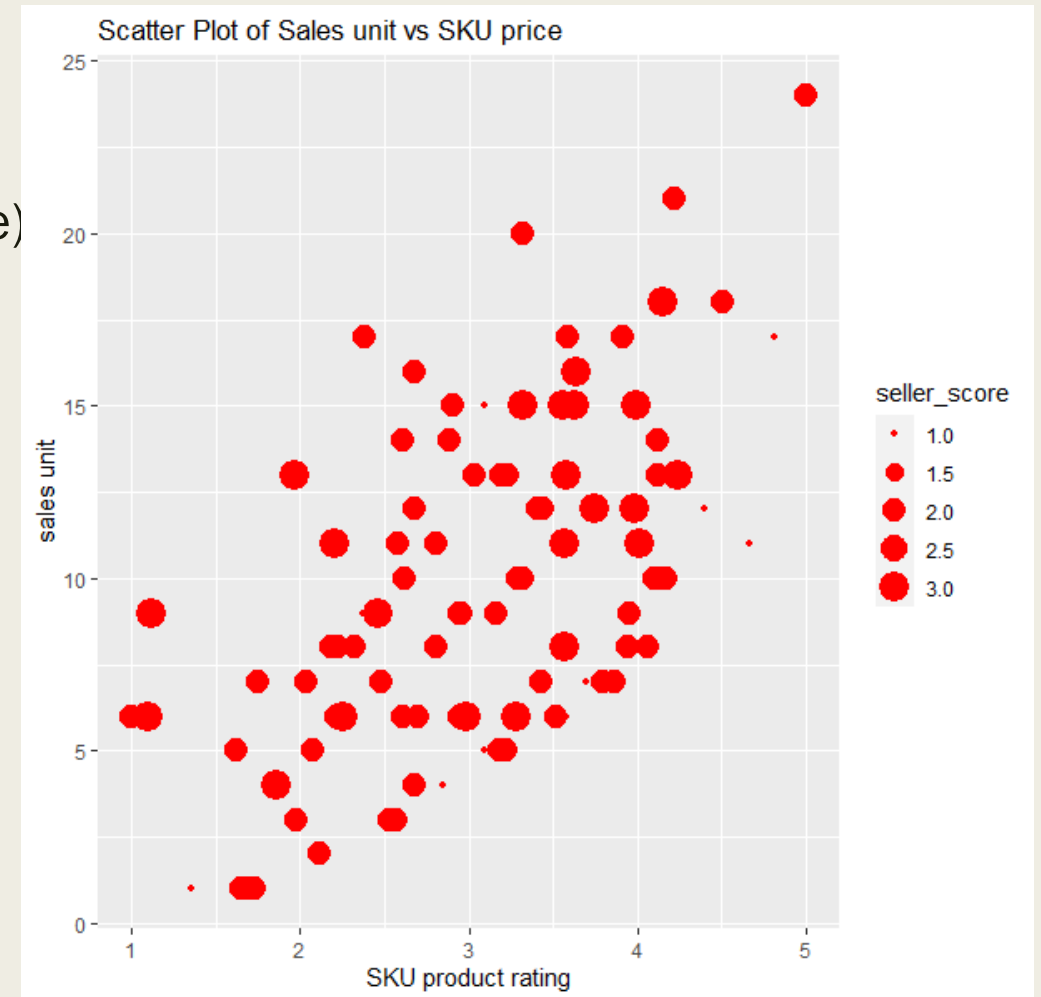
- Continuous variable “price” is mapped to color.

```
ggplot(sales_jul[1:100,])+  
  aes(x=prod_rating,y=sales_unit,col=price)+  
  geom_point()+  
  labs(y="sales unit",  
        x="SKU product rating",  
        title = "Scatter Plot of Sales unit vs SKU price")
```



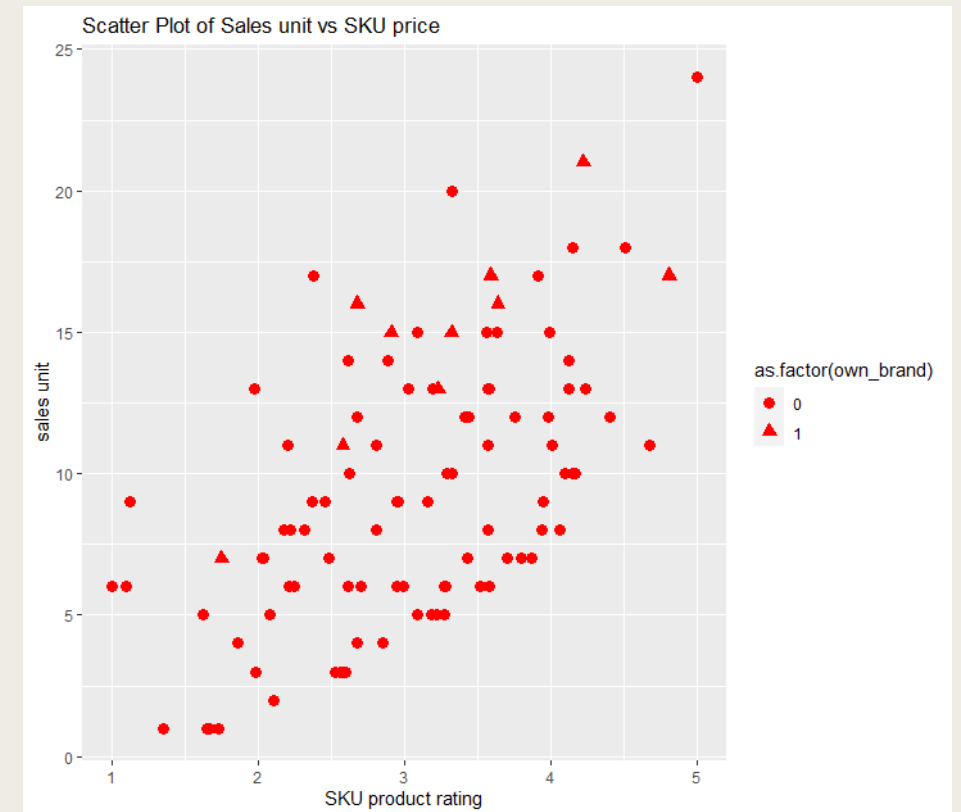
Scatter plot with mapping size

```
ggplot(sales_jul[1:100,])+  
  aes(x=prod_rating,y=sales_unit,size = seller_score)+  
  geom_point(col="red")+  
  labs(y="sales unit",  
        x="SKU product rating",  
        title = "Scatter Plot of Sales unit vs SKU price")
```



Scatter plot with mapping shape

```
ggplot(sales_jul[1:100,])+  
  aes(x=prod_rating,y=sales_unit,shape =  
as.factor(own_brand))+  
  geom_point(col="red",size=3)+  
  labs(y="sales unit",  
        x="SKU product rating",  
        title = "Scatter Plot of Sales unit vs SKU price")
```



Combination

```
ggplot(sales_jul[1:100,])+  
  aes(x=prod_rating,y=sales_unit,col=as.factor(prom),shape =  
as.factor(own_brand))+  
  geom_point(size = 3)+  
  labs(y="sales unit",  
        x="SKU product rating",  
        title = "Scatter Plot of Sales unit vs SKU price")
```



Add smooth line

- `geom_smooth()` uses `method = "loess"` for small number of observations and formula `"y ~ x"`

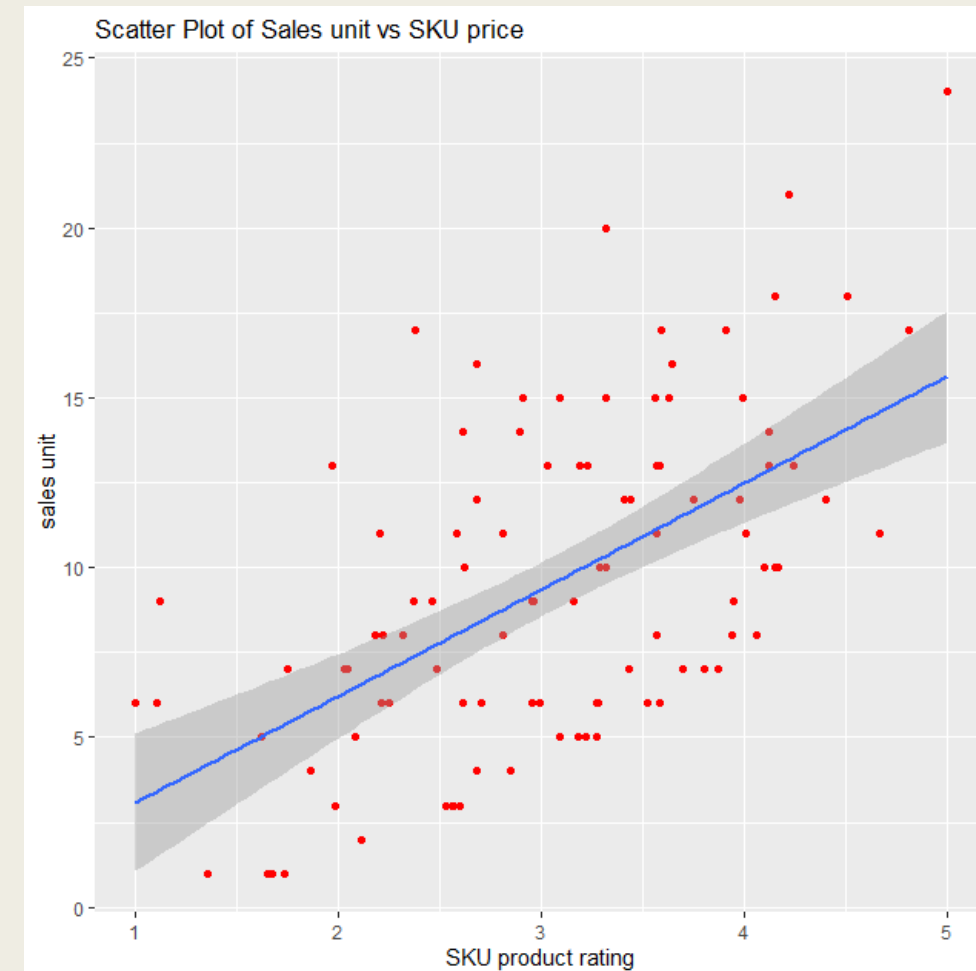
```
ggplot(sales_jul[1:100,])+  
  aes(x=prod_rating,y=sales_unit)+  
  geom_point(col="red")+  
  geom_smooth()+  
  labs(y="sales unit",  
       x="SKU product rating",  
       title = "Scatter Plot of Sales unit vs SKU price")
```



Add linear smoother

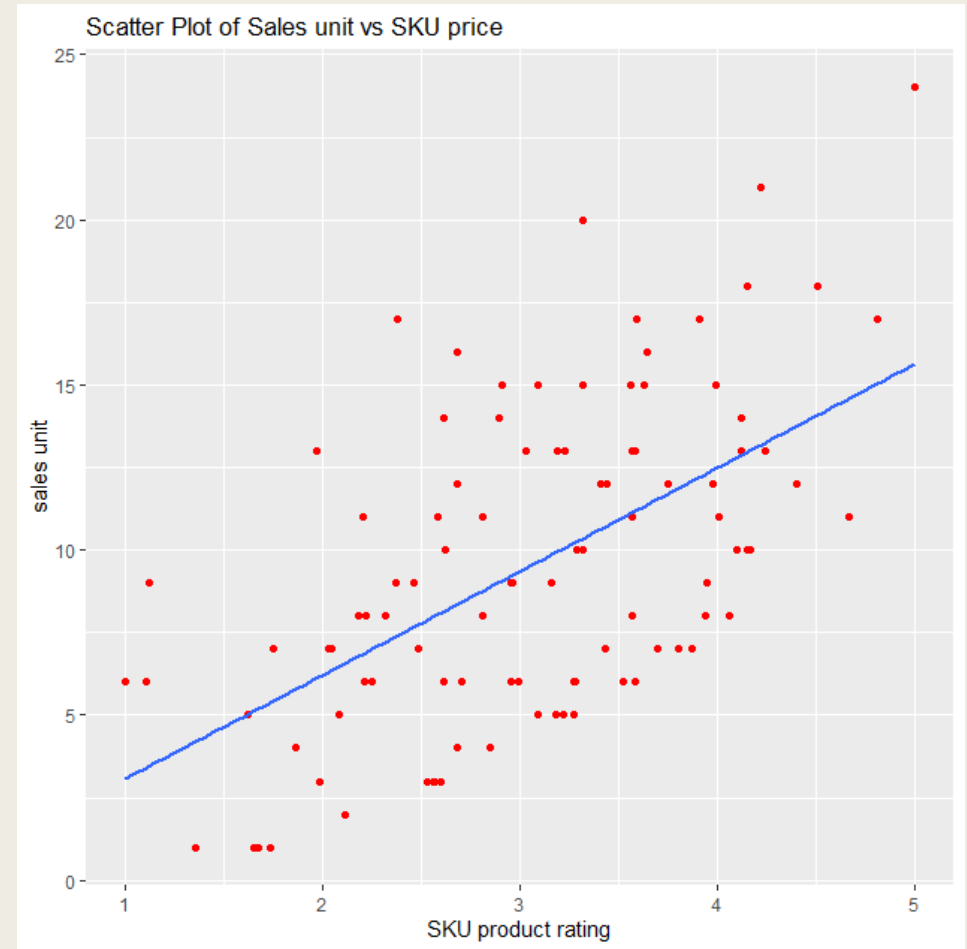
- There are many different smoothers you can choose between by using "method" argument.

```
ggplot(sales_jul[1:100,])+  
  aes(x=prod_rating,y=sales_unit)+  
  geom_point(col="red")+  
  geom_smooth(method = "lm")+  
  labs(y="sales unit",  
       x="SKU product rating",  
       title = "Scatter Plot of Sales unit vs SKU price")
```



Add smoother without confidence interval

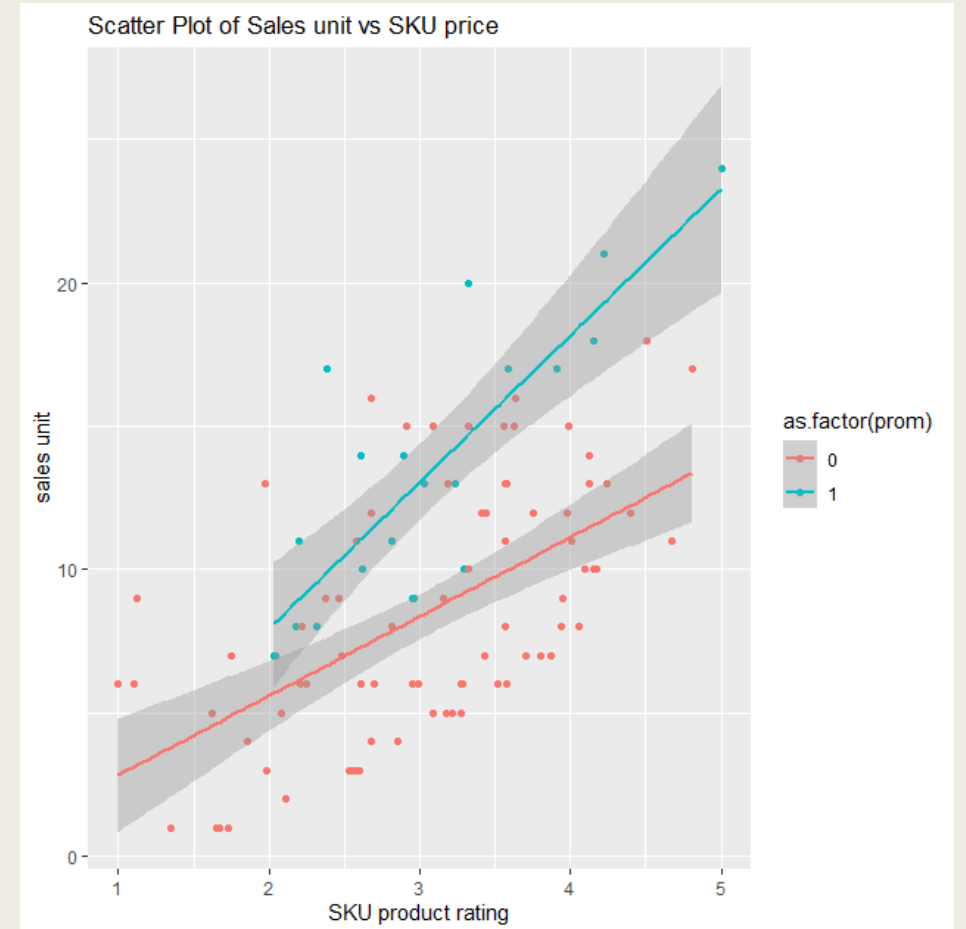
```
ggplot(sales_jul[1:100,])+  
  aes(x=prod_rating,y=sales_unit)+  
  geom_point(col="red")+  
  geom_smooth(method = "lm",se = FALSE)+  
  labs(y="sales unit",  
       x="SKU product rating",  
       title = "Scatter Plot of Sales unit vs SKU price")
```



Add smoothers by group

Smoothers are automatically t to each group

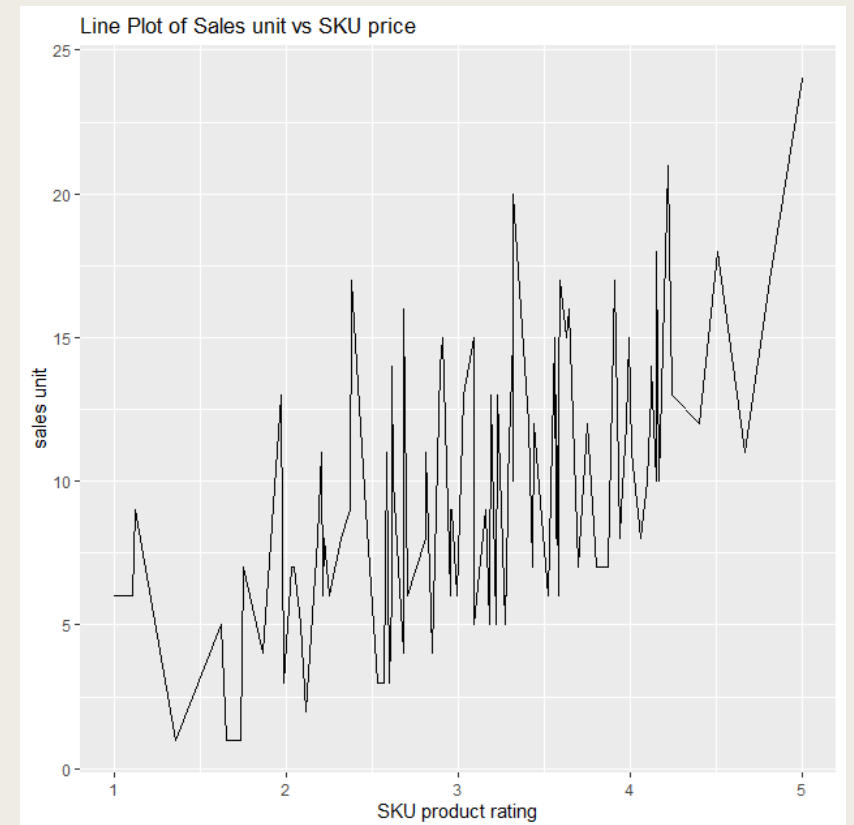
```
ggplot(sales_jul[1:100,])+  
  aes(x=prod_rating,y=sales_unit,col = as.factor(prom))+  
  geom_point()+  
  geom_smooth(method = "lm")+  
  labs(y="sales unit",  
       x="SKU product rating",  
       title = "Scatter Plot of Sales unit vs SKU price")
```



Line plot

- `geom_line()` connects points in order of the variable on the x axis. It is suitable for time series.

```
ggplot(sales_jul[1:100,])+  
  aes(x=prod_rating,y=sales_unit)+  
  geom_line()+  
  labs(y="sales unit",  
        x="SKU product rating",  
        title = "Line Plot of Sales unit vs SKU price")
```



Correlation coefficient to measure linear relationship

- When **a relationship between variables appears to be linear**, the correlation provides a numeric measure of the strength of that relationship.
 - The correlation is always a number $-1 \leq r \leq 1$,
 - *r-values close to -1 or $+1$ indicate a strong relationship,*
 - *the closer r is to zero the less strength in the linear relationship,*
 - *negative/positive values go with negative/positive associations*
 - In R, cor() command
- ```
> cor(sales_jul$prod_rating, sales_jul$sales_unit)
```
- ```
[1] 0.3931626
```

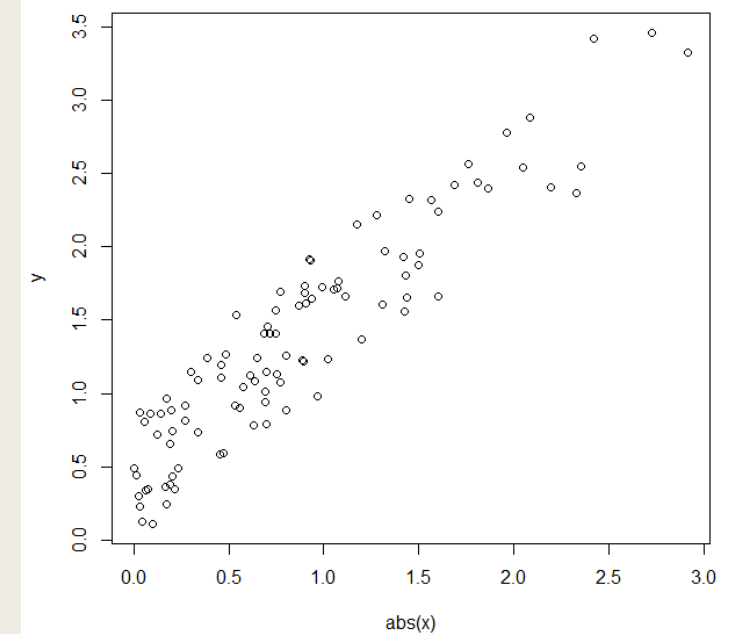
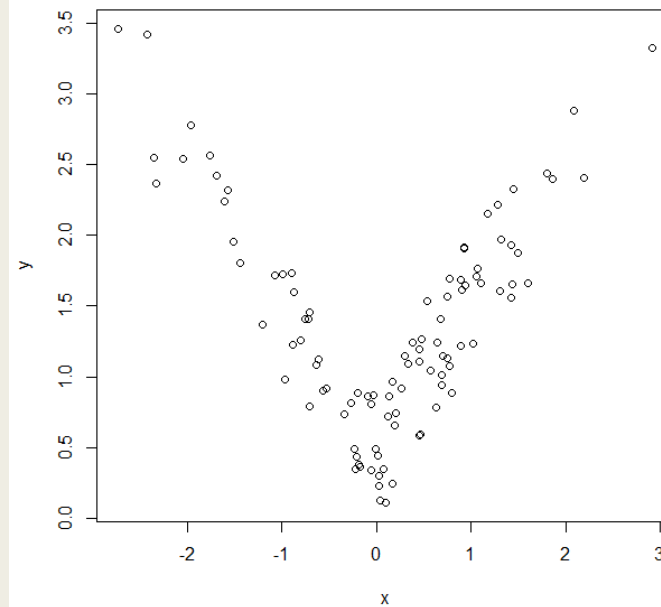
Does not work for non-linear relation

```
x=rnorm(100)
```

```
y=abs(x)+runif(100)
```

```
cor(x,y) -> -0.07773027
```

```
cor(abs(x),y) -> 0.9316618
```



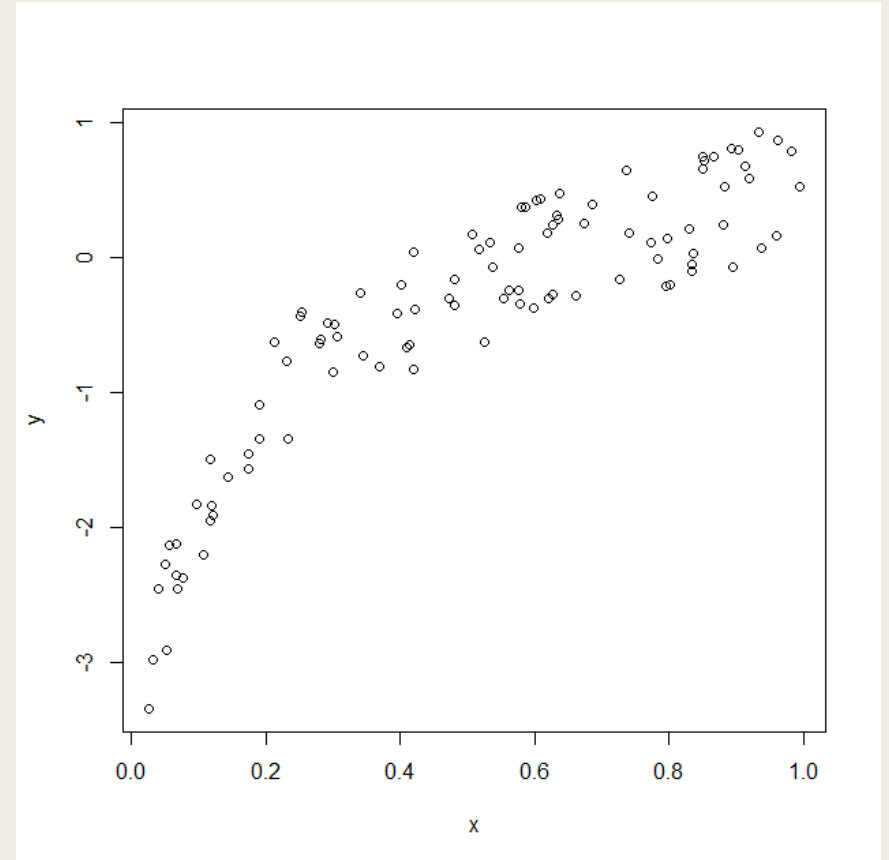
Check relation for transformation

```
x<-runif(100)
```

```
y<-log(x)+runif(100)
```

```
cor(x,y) -> 0.88
```

```
cor(log(x),y) → 0.94
```



Linear regression

```
lm_1<-lm(sales_unit~prod_rating, data =  
sales_jul)
```

```
summary(lm_1) #testing result
```

Call:

```
lm(formula = sales_unit ~ prod_rating, data = sales_jul)
```

Residuals:

Min	1Q	Median	3Q	Max
-10.0727	-4.5736	-0.2044	4.3409	15.4421

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.8213	0.8254	7.053	4.78e-12 ***
prod_rating	2.7505	0.2607	10.552	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.014 on 609 degrees of freedom
Multiple R-squared: 0.1546, Adjusted R-squared: 0.1532
F-statistic: 111.3 on 1 and 609 DF, p-value: < 2.2e-16

The model above is achieved by using the `lm()` function in R and the output is called using the `summary()` function on the model.

How to interpret linear regression results in R

- Residual: Residuals are essentially the difference between the actual observed response values and the response values that the model predicted.

Residuals:

Min	1Q	Median	3Q	Max
-10.0727	-4.5736	-0.2044	4.3409	15.4421

Two-sided T-test in linear regression model

Use a linear regression to determine whether the slope of the regression line differs significantly from zero.

$H_0: B_1 = 0$ vs $H_a: B_1 \neq 0$

- Coefficients:
- Estimate Std. Error t value Pr(> |t|)
- (Intercept) 5.8213 0.8254 7.053 4.78e-12 ***
- prod_rating 2.7505 0.2607 10.552 < 2e-16 ***

DF = n - 2

Interpret results. Since the P-value is less than the significance level (0.05), we cannot accept the null hypothesis.

F-test in linear regression model

■ Multiple R-squared, Adjusted R-squared

The **R-squared (R²) statistic** provides a measure of how well the model is fitting the actual data. It takes the form of a proportion of variance. R² is a measure of the linear relationship between our predictor variable and our response / target variable

Multiple R-squared: 0.1546, Adjusted R-squared: 0.1532

15.46% variance of dependent variable has been explained by the predictor

In multiple regression settings, the R² will always increase as more variables are included in the model. That's why the **adjusted R²** is the preferred measure as it adjusts for the number of variables considered.

F-test

$$F = \frac{\text{explained variation}/(k-1)}{\text{unexplained variation}/(n-k)}$$
$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

Where, k is the number of variable + 1;
n is the sample size;
DF1 = k-1; DF2=n-k

F-statistics

- it's hard to define what level of R^2 is appropriate to claim the model fits well.
- T-test: judging coefficients of individual variables on their own for significance
- F statistic: judges on multiple coefficients taken together at the same time. F – Test for overall significance compares an intercept only regression model with the current model.
- *H0 : The fit of intercept only model and the current model is same. i.e. Additional variables do not provide value taken together*

Ha : The fit of intercept only model is significantly less compared to our current model. i.e. Additional variables do make the model significantly better.