R Assignment - Group 04

DC Character Debut by Year (2010-2020)

Table of Content

1)Introduction of the Analysis

This dataset contains information about Marvel and DC characters from 1939 until 2014 (August 24th). It was used for the 538 study. This Data set includes-following sections.

- Year: Year of First Appearance
- Character: Name of Character
- Character-href: URL leading to detailed info page of each character
- Real Name: Real name of the character, if present
- Current Alias: Commonly known name/identity
- Alignment: Whether the character is good, bad or neutral
- Identity: Whether the alias is public or secret
- Citizenship: Citizenship of the character, if present
- Marital Status: Whether the character is married or single
- Occupation: Normal occupation of the character, if present
- Gender: Gender of character
- Hair: Hair color
- Eye: Eye color
- Universe: To which universe the character belongs to
- First Appearance: Exact comic, volume and date (to month) where the character first appears
- The appearance of Death: Exact comic, volume and date (to month) where the character dies, if present

2)The observation about the data set

I) Alignment and Identity

- In this table we are considering each Dc Characters Alignment & thier identity for analyzation

We have used following libraries:

>library(ggplot2)

>library(dplyr)

>library(gridExtra)

>library(RColorBrewer)

>library(wordcloud)

>library(plotrix)

>library(fmsb)

>library(fivethirtyeight)

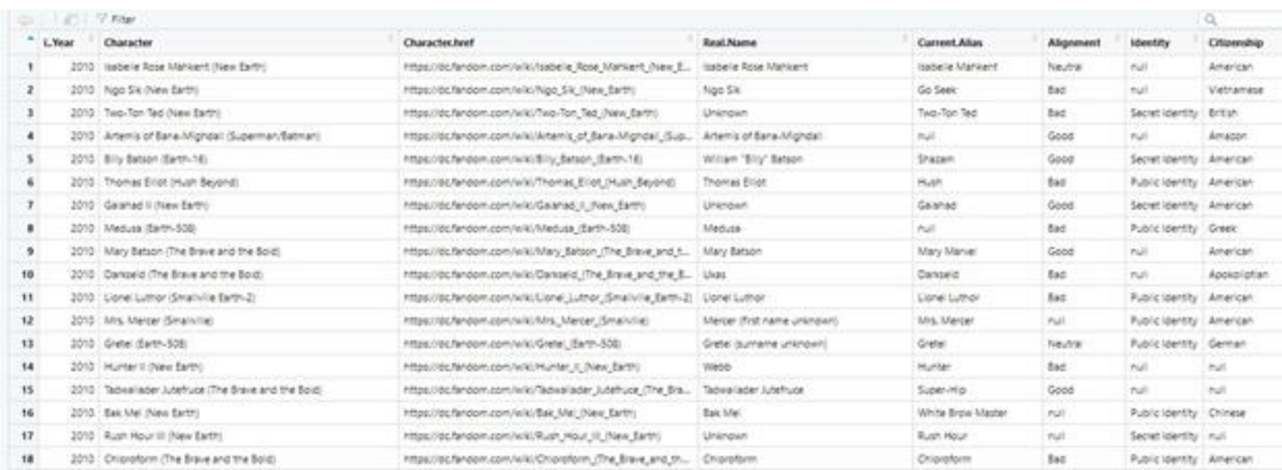
>library(knitr)

Read the .csv file and group the selected column variables

> DC<-read.csv("dc_2010_2020.csv")

Display the table of the selected data set.

> View(DC)



	L.Year	Character	Character.href	Real.Name	Current.Alias	Alignment	Identity	Citizenship
1	2010	Isabelle Rose Markert (New Earth)	https://dc.fandom.com/wiki/Isabelle_Rose_Markert_(New_E...	Isabelle Rose Markert	Isabelle Markert	Neutral	null	American
2	2010	Ngo Sik (New Earth)	https://dc.fandom.com/wiki/Ngo_Sik_(New_Earth)	Ngo Sik	Go Seek	Bad	null	Vietnamese
3	2010	Two-Ton Ted (New Earth)	https://dc.fandom.com/wiki/Two-Ton_Ted_(New_Earth)	Unknown	Two-Ton Ted	Bad	Secret Identity	British
4	2010	Artemis of Bana-Mighdall (Superman/Batman)	https://dc.fandom.com/wiki/Artemis_of_Bana-Mighdall_(Sup...	Artemis of Bana-Mighdall	null	Good	null	Amazon
5	2010	Billy Batson (Earth-16)	https://dc.fandom.com/wiki/Billy_Batson_(Earth-16)	William "Billy" Batson	Shazam	Good	Secret Identity	American
6	2010	Thomas Elliot (Hush Beyond)	https://dc.fandom.com/wiki/Thomas_Elliot_(Hush_Beyond)	Thomas Elliot	Hush	Bad	Public Identity	American
7	2010	Galahad II (New Earth)	https://dc.fandom.com/wiki/Galahad_II_(New_Earth)	Unknown	Galahad	Good	Secret Identity	American
8	2010	Medusa (Earth-508)	https://dc.fandom.com/wiki/Medusa_(Earth-508)	Medusa	null	Bad	Public Identity	Greek
9	2010	Mary Batson (The Brave and the Bold)	https://dc.fandom.com/wiki/Mary_Batson_(The_Brave_and_t...	Mary Batson	Mary Marvel	Good	null	American
10	2010	Darkseid (The Brave and the Bold)	https://dc.fandom.com/wiki/Darkseid_(The_Brave_and_the_B...	Uxas	Darkseid	Bad	null	Apokoliptian
11	2010	Lionel Luthor (Smallville Earth-2)	https://dc.fandom.com/wiki/Lionel_Luthor_(Smallville_Earth-2)	Lionel Luthor	Lionel Luthor	Bad	Public Identity	American
12	2010	Mrs. Mercer (Smallville)	https://dc.fandom.com/wiki/Mrs._Mercer_(Smallville)	Mercer (first name unknown)	Mrs. Mercer	null	Public Identity	American
13	2010	Gretel (Earth-508)	https://dc.fandom.com/wiki/Gretel_(Earth-508)	Gretel (surname unknown)	Gretel	Neutral	Public Identity	German
14	2010	Hunter II (New Earth)	https://dc.fandom.com/wiki/Hunter_II_(New_Earth)	Webb	Hunter	Bad	null	null
15	2010	Tadwallader Jutefruce (The Brave and the Bold)	https://dc.fandom.com/wiki/Tadwallader_Jutefruce_(The_Bra...	Tadwallader Jutefruce	Super-Hip	Good	null	null
16	2010	Bak Mei (New Earth)	https://dc.fandom.com/wiki/Bak_Mei_(New_Earth)	Bak Mei	White Brow Master	null	Public Identity	Chinese
17	2010	Rush Hour III (New Earth)	https://dc.fandom.com/wiki/Rush_Hour_III_(New_Earth)	Unknown	Rush Hour	null	Secret Identity	null
18	2010	Chloroform (The Brave and the Bold)	https://dc.fandom.com/wiki/Chloroform_(The_Brave_and_th...	Chloroform	Chloroform	Bad	Public Identity	American

2.1 Original Dataset In-view

Summarize the details from original Dataset.

> DCAI<-DC %>% group_by(Alignment,Identity) %>% summarise(number = n()) %>%

arrange(-number)

2.2 Table view of the above summarized data.

	Alignment	Identity	number
1	Good	Public Identity	2336
2	Good	Secret Identity	1542
3	Bad	Public Identity	1405
4	Bad	Secret Identity	1389
5	Neutral	Public Identity	694
6	Good	null	472
7	Bad	null	460
8	Neutral	Secret Identity	319
9	null	Public Identity	148
10	null	null	139
11	Neutral	null	106
12	null	Secret Identity	94
13			11
14	Bad	Secret	2
15	Good	Public	1
16	Good	public Identity	1

- Group the Alignment and Identity, then calculate the total identity count with each alignment category. Then add the percentage label.

> DU<-DCAI %>% group_by(Identity) %>% mutate(countT= sum(number)) %>% group_by(Alignment)   %>% mutate(percentage=100*number/countT)

- Add a percentage label and round up values to two decimal places.

> DU$LABEL <-paste0(round(DU$percentage,2))

	Alignment	Identity	number	countT	percentage	LABEL
1	Good	Public Identity	2336	4583	50.970980	50.97
2	Good	Secret Identity	1542	3344	46.112440	46.11
3	Bad	Public Identity	1405	4583	30.656775	30.66
4	Bad	Secret Identity	1389	3344	41.537081	41.54
5	Neutral	Public Identity	694	4583	15.142919	15.14
6	Good	null	472	1177	40.101954	40.1
7	Bad	null	460	1177	39.082413	39.08
8	Neutral	Secret Identity	319	3344	9.539474	9.54
9	null	Public Identity	148	4583	3.229326	3.23
10	null	null	139	1177	11.809686	11.81
11	Neutral	null	106	1177	9.005947	9.01
12	null	Secret Identity	94	3344	2.811005	2.81
13			11	11	100.000000	100
14	Bad	Secret	2	2	100.000000	100
15	Good	Public	1	1	100.000000	100
16	Good	public Identity	1	1	100.000000	100

2.3 Filtered Data Grouped with Identity & Alignment

pieC<-as.data.frame(DCAI %>% group_by(Identity) %>% select(number) %>% summarise(sum=sum(number)))

Name	Type	Value
g1	list [9] (S3: gg, ggplot)	List of length 9
data	list [16 x 7] (S3: grouped_df, tbl_	A tibble with 16 rows and 7 columns
layers	list [2]	List of length 2
scales	environment [2] (S3: ScalesList, g	<environment: 0x000001336115afe8>
mapping	list [3] (S3: uneval)	List of length 3
theme	list [2]	List of length 2
coordinates	environment [5] (S3: CoordCarte	<environment: 0x0000013361a33828>
facet	environment [2] (S3: FacetNull, F	<environment: 0x00000133619d7930>
plot_env	environment [7]	<environment: R_GlobalEnv>
labels	list [4]	List of length 4

- Plotting above summarizes data with ggplot2.

> g1<-ggplot(data=DU,aes(x=Alignment,y=percentage,fill=Identity)) + geom_bar(width = 0.9, stat="identity",position='dodge') + theme(axis.text.x = element_text(angle=90,

hjust=1),legend.position='none') + geom_text(aes(label=LABEL),

position=position_dodge(width=0.9), vjust=-0.25,size=2.5)

+ scale_fill_manual(values

=c("olivedrab","steelblue","red","yellow","green","black","orange")) + xlab('') +

ylab('Percentage')+ scale_colour_manual("",breaks = c("Unknown", "null", "Public

Identity","Secret Identity","Secret","Public","public Identity"), values = c("olivedrab"

, "steelblue", "green","orange","black","red","yellow"))

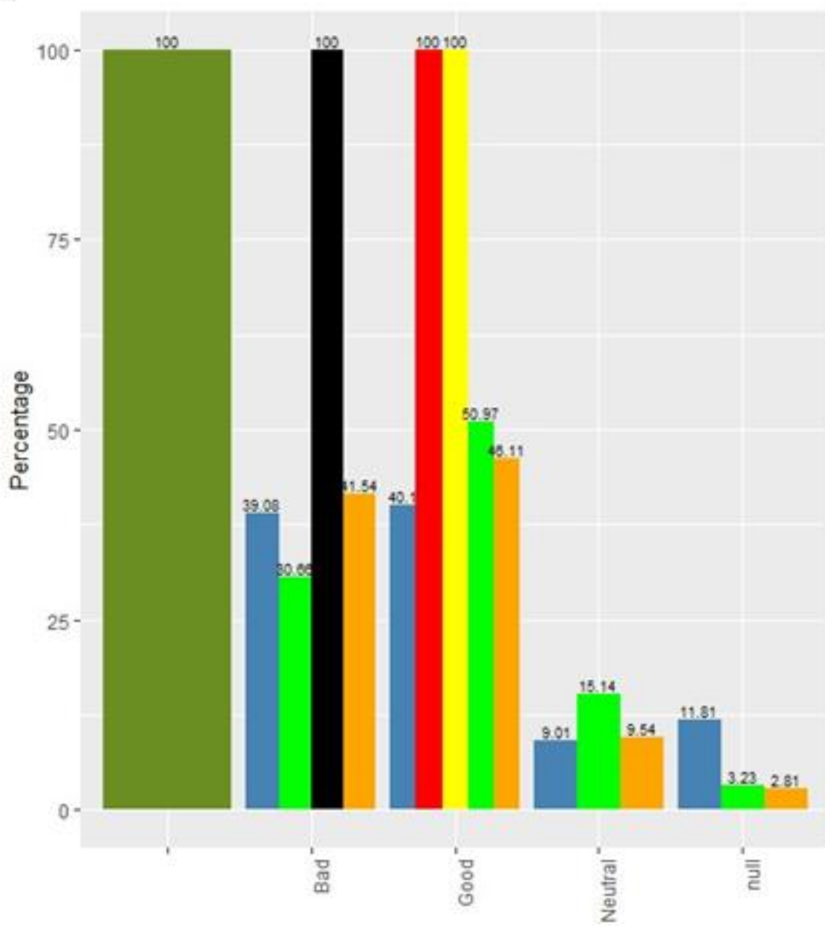
Plot Details for Data observation Respectively

Extra plot Details

- **Olive-drab**: Unknown
- **Steelblue**:null
- **Red:** Public Identity
- **Yellow:** Secret Identity
- **Green:** Secret
- **Black:** Public
- **Orange:** public Identity

	Alignment	Identity	number	countT	percentage	LABEL
1	Good	Public Identity	2336	4583	50.970980	50.97
2	Good	Secret Identity	1542	3344	46.112440	46.11
3	Bad	Public Identity	1405	4583	30.656775	30.66
4	Bad	Secret Identity	1389	3344	41.537081	41.54
5	Neutral	Public Identity	694	4583	15.142919	15.14
6	Good	null	472	1177	40.101954	40.1
7	Bad	null	460	1177	39.082413	39.08
8	Neutral	Secret Identity	319	3344	9.539474	9.54
9	null	Public Identity	148	4583	3.229326	3.23
10	null	null	139	1177	11.809686	11.81
11	Neutral	null	106	1177	9.005947	9.01
12	null	Secret Identity	94	3344	2.811005	2.81
13			11	11	100.000000	100
14	Bad	Secret	2	2	100.000000	100
15	Good	Public	1	1	100.000000	100
16	Good	public Identity	1	1	100.000000	100

2.4 Filtered Data Grouped with Identity & Alignment Table



2.5 Filtered Data representation in a plot Alignment-wise with respective to identity for each Alignment (Identity against Alignment Plot)

II) Alignment Frequency

- We get Alignment into the X-axis and display each identity as a percentage.

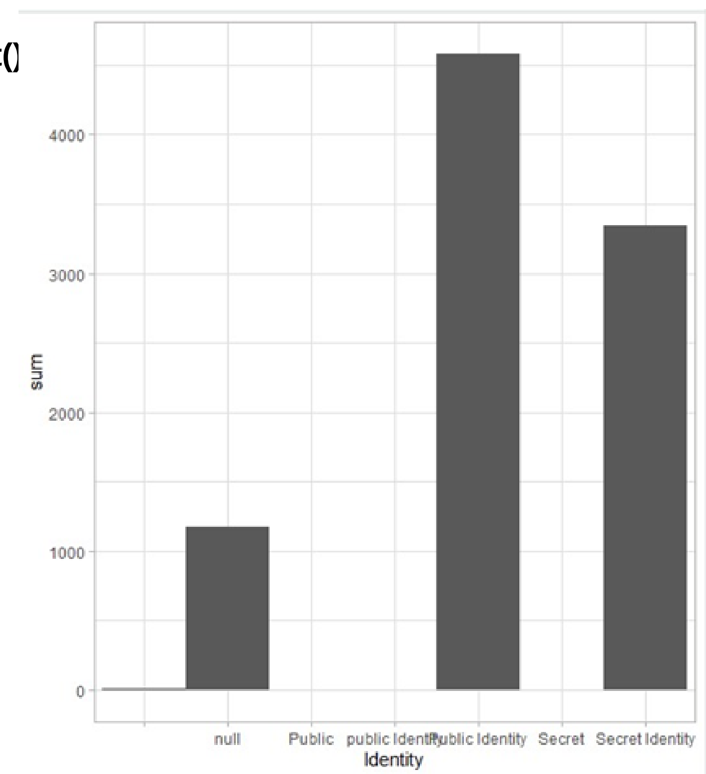
```
> g2<-ggplot(pieC,aes(x="",y=sum,fill=Identity)) + geom_bar(stat='identity',width =
1) +coord_polar(theta="y") + theme_void()
+theme(axis.text.x=element_blank(),legend.position='bottom')
+scale_fill_manual(values
=c("olivedrab","steelblue","red","yellow","green","black","orange"))
+geom_text(aes(y =c(20000,8000), label = paste(pieC$Alignment,": ",pieC$sum)))
```

```
> p1<-ggplot(data = pieC,aes(x=Identity,y= sum))

> p1<-p1+ geom_bar(width=2 , stat = "identity")

> p1<-p1+ theme(legend.position = "none")

> p1<-p1+ theme_light()
```



2.6 (This Plot represents the Counts of each Identity category)

III) Alignment & Marital Status Representation

- Creating a Table to Filter Marital status and Identity

DL<-DC %>% group_by(Marital.Status,Identity) %>% summarise(number = n()) %>% arrange(-number)

	Marital.Status	Identity	number
1	null	Public Identity	1955
2	Single	Public Identity	1925
3	Single	Secret Identity	1629
4	null	Secret Identity	1473
5	null	null	882
6	Married	Public Identity	435
7	Single	null	234
8	Widowed	Public Identity	157
9	Married	Secret Identity	117
10	Widowed	Secret Identity	80
11	Divorced	Public Identity	71
12	Married	null	40
13	Divorced	Secret Identity	20
14	Engaged	Public Identity	19
15	Separated	Public Identity	18
16	Engaged	Secret Identity	16
17			11
18	Widowed	null	10
19	Separated	Secret Identity	7
20	Divorced	null	5
21	Engaged	null	3
22	null	Secret	2
23	Separated	null	2
24	Divorced Widowed	null	1
25	Married Divorced	Public Identity	1
26	null	Public	1
27	null	public Identity	1
28	Remarried	Public Identity	1
29	Remarried	Secret Identity	1
30	Widowed Married	Public Identity	1
31	Widowed Single	Secret Identity	1

2.7 Table represents the Counts of each Identity with Martial Status

IV) Gender & Marital Status Representation

- Creating a Table to Filter Marital status and Gender with Gender-wise segregation of Martial Status. Percentage denoted that the basing Gender, how the Characters marital status separately and given it as an precentage

DN<-DC %>% group_by(Marital.Status,Gender) %>% summarise(number = n()) %>% arrange(-number)

DNN<-DN%>% group_by(Gender) %>% mutate(countT= sum(number)) %>% group_by(Marital.Status) %>% mutate(percentage=100*number/countT)

	Marital.Status	Gender	number	countT	percentage
1	null	Male	3180	6102	52.11406096
2	Single	Male	2314	6102	37.92199279
3	Single	Female	1420	2840	50.00000000
4	null	Female	1023	2840	36.02112676
5	Married	Male	344	6102	5.63749590
6	Married	Female	247	2840	8.69718310
7	Widowed	Male	165	6102	2.70403147
8	null	null	95	126	75.39682540
9	Widowed	Female	82	2840	2.88732394
10	Divorced	Male	62	6102	1.01606031
11	Divorced	Female	34	2840	1.19718310
12	Single	null	30	126	23.80952381
13	Engaged	Female	19	2840	0.66901408
14	Engaged	Male	19	6102	0.31137332
15	null	Genderless	16	31	51.61290323
16	Separated	Male	16	6102	0.26220911
17	Single	Genderless	15	31	48.38709677
18			11	11	100.00000000
19	Separated	Female	11	2840	0.38732394
20	Single	Transgender	6	6	100.00000000
21	Single	Non-binary	3	3	100.00000000
22	Divorced Widowed	Female	1	2840	0.03521127
23	Married	null	1	126	0.79365079
24	Married Divorced	Male	1	6102	0.01638807
25	Remarried	Female	1	2840	0.03521127
26	Remarried	Male	1	6102	0.01638807
27	Widowed Married	Female	1	2840	0.03521127
28	Widowed Single	Female	1	2840	0.03521127

2.8 Table represents the Counts of each Gender with Martial Status

3) Appropriate Plots/Charts

> DC[1,]

- Getting row by row details.

```
> DC[1,]
  i..Year                 Character                                       Character.href              Real.Name  Current.Alias Alignment
1    2010 Isabelle Rose Mahkent (New Earth) https://dc.fandom.com/wiki/Isabelle_Rose_Mahkent_(New_Earth) Isabelle Rose Mahkent Isabelle Mahkent   Neutral
  Identity Citizenship Marital.Status Occupation Gender Hair Eyes Universe                 First.Appearance Appearance.of.Death
1     null    American         Single       null Female null null     null JSA All-Stars  #11\n(December, 2010)                null
> |
```

- Filtering the Alignment Levels.

> DCNEW <- droplevels(filter(DC,Alignment != "null"))

> head(DCNEW)

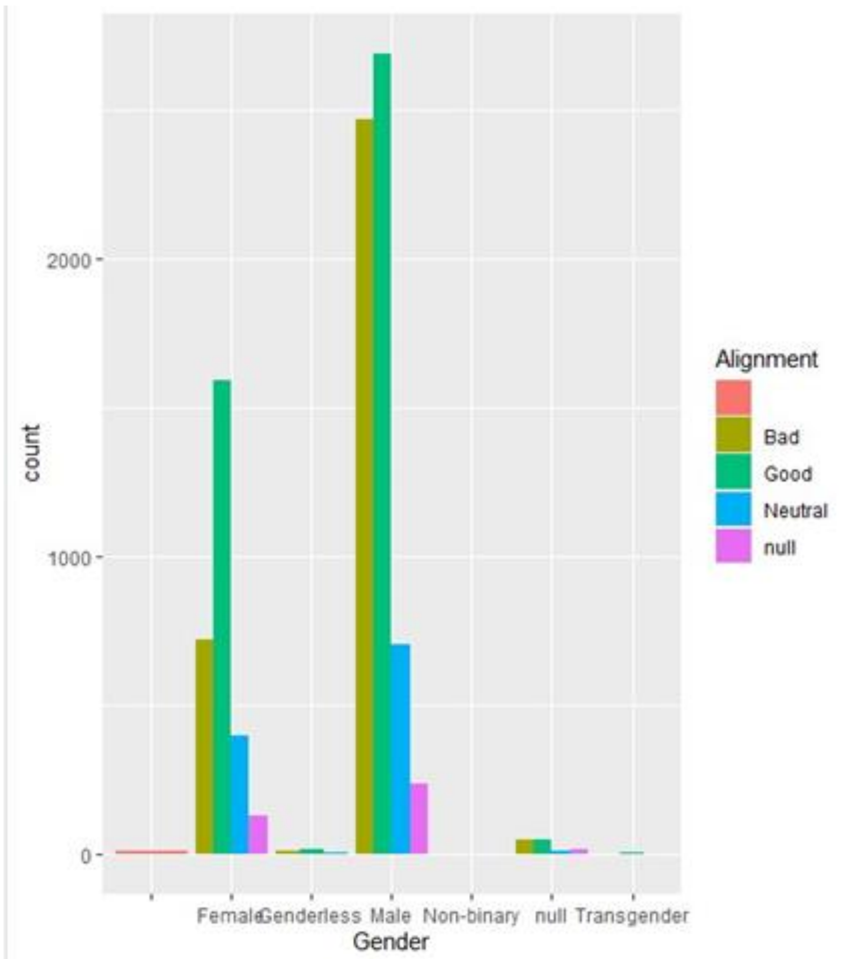
```
> DCNEW <- droplevels(filter(DC,Alignment != "null"))
> head(DCNEW)
  i..Year                            Character                                                    Character.href                 Real.Name
1    2010            Isabelle Rose Mahkent (New Earth)          https://dc.fandom.com/wiki/Isabelle_Rose_Mahkent_(New_Earth)     Isabelle Rose Mahkent
2    2010                        Ngo Sik (New Earth)                        https://dc.fandom.com/wiki/Ngo_Sik_(New_Earth)                   Ngo Sik
3    2010                   Two-Ton Ted (New Earth)                    https://dc.fandom.com/wiki/Two-Ton_Ted_(New_Earth)                   Unknown
4    2010 Artemis of Bana-Mighdall (Superman/Batman) https://dc.fandom.com/wiki/Artemis_of_Bana-Mighdall_(Superman/Batman) Artemis of Bana-Mighdall
5    2010                  Billy Batson (Earth-16)                    https://dc.fandom.com/wiki/Billy_Batson_(Earth-16)   William "Billy" Batson
6    2010                Thomas Elliot (Hush Beyond)                https://dc.fandom.com/wiki/Thomas_Elliot_(Hush_Beyond)             Thomas Elliot
     Current.Alias Alignment        Identity Citizenship Marital.Status Occupation Gender  Hair  Eyes                 Universe
1 Isabelle Mahkent   Neutral            null    American         Single       null Female  null  null                     null
2          Go Seek       Bad            null  Vietnamese           null  Kidnapper   Male  null  null                New Earth
3      Two-Ton Ted       Bad Secret Identity     British           null       null   Male  null  null                New Earth
4             null      Good            null      Amazon         Single       null Female   Red Green Superman/Batman (Reality)
5           Shazam      Good Secret Identity    American         Single Adventurer   Male Black  Blue                 Earth-16
6             Hush       Bad Public Identity    American         Single    Surgeon   Male   Red  Blue              Hush Beyond
                    First.Appearance               Appearance.of.Death
1       JSA All-Stars  #11\n(December, 2010)                      null
2            Azrael Vol 2  #7\n(June, 2010)        Azrael Vol 2  #7\n(June, 2010)
3   Knight and Squire  #1\n(December, 2010)                      null
4                                  null                              null
5                                  null                              null
6 Batman Beyond Vol 3  #2\n(September, 2010) Batman Beyond Vol 3  #5\n(December, 2010)
> |
```

3.1 Data Segregation for Alignment Analysis

I) Create a side-by-side bar chart of gender by Align variable using ggplot2 for Data Observation

> ggplot(DC, aes(x = Gender, fill =Alignment)) + geom_bar(position = "dodge")

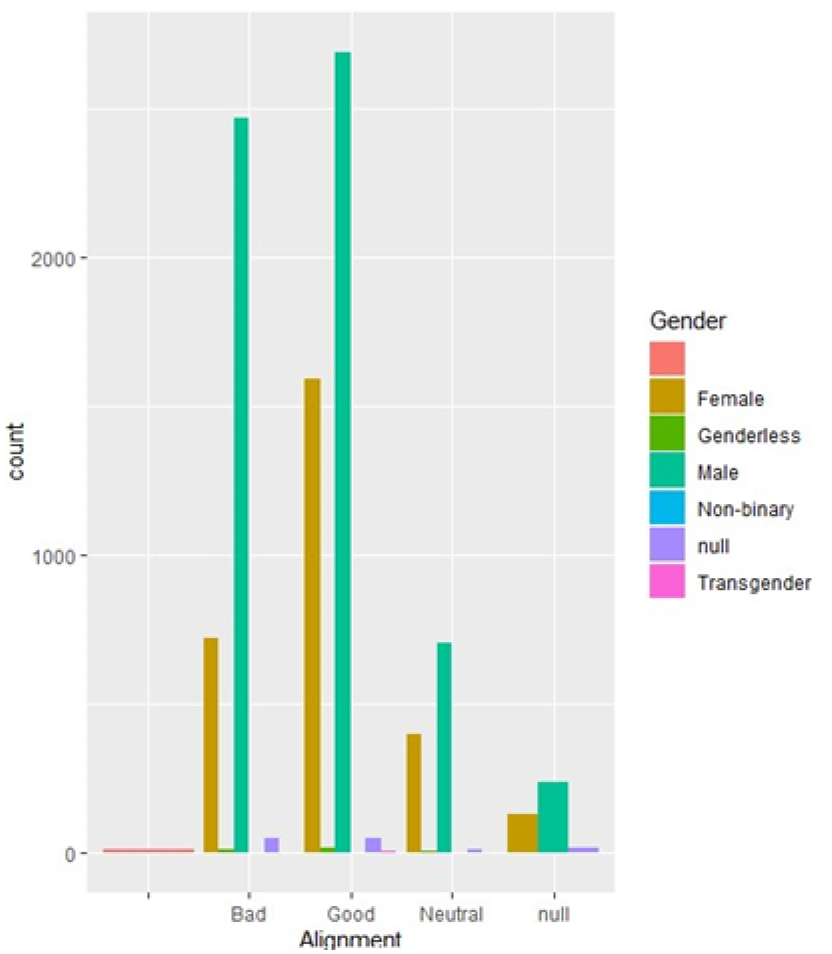
```
> ggplot(DC, aes(x = Identity, fill = Alignment )) + geom_bar(position = "fill")
> |
```



3.2 (This above plot represent the Gender count with the Alignment category. In X-axis we get Gender categories and Y-Axis displayed each count in that gender category.)

> ggplot(DC, aes(x = Alignment, fill = Gender)) + geom_bar(position = "dodge")

```
> ggplot(DC, aes(x = Alignment, fill = Identity )) + geom_bar(position = "fill")
>
```



3.3 - (This above plot represent the Alignment count with the Gender category. In X-axis we get Alignment categories and Y-Axis displayed each count in that Alignment category.)

<u>II) Proportion diagrams for Data Observation</u>

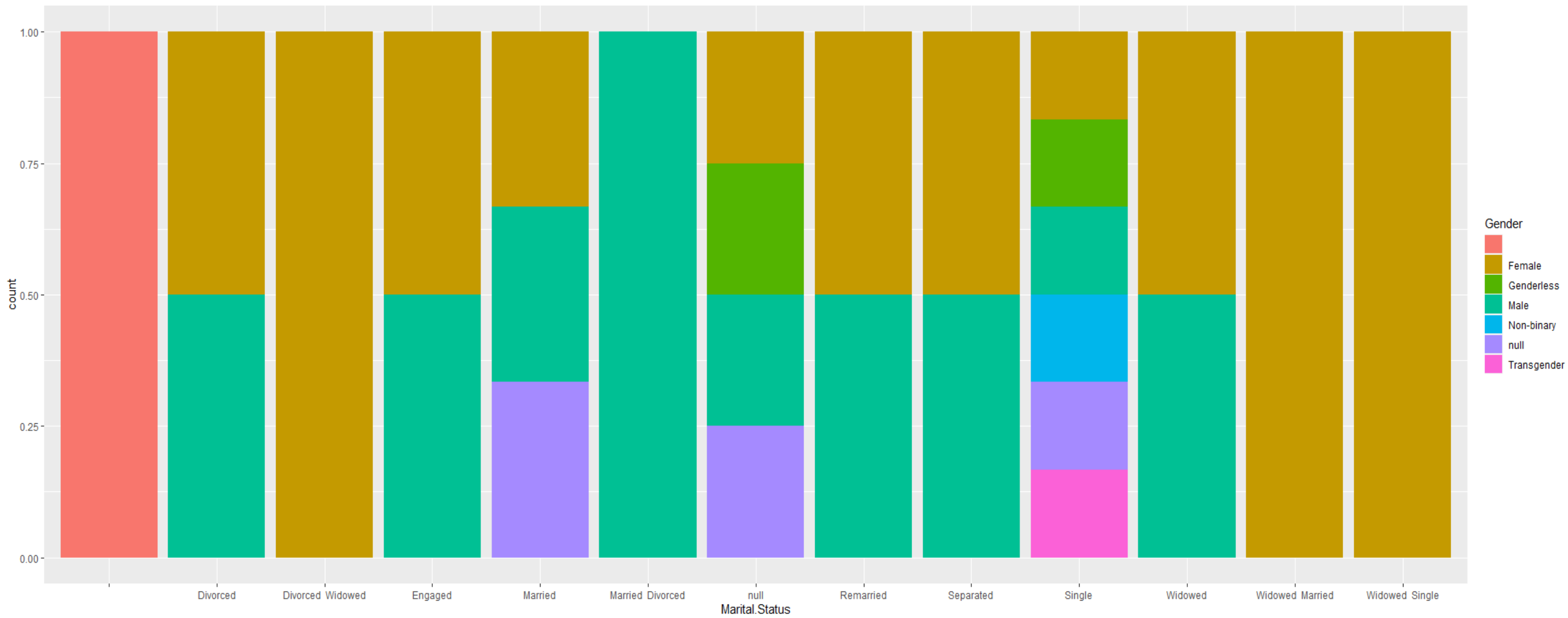
1) <u>Plot Marital Status against Gender Variation of Each Character</u>

> DN<-DC %>% group_by(Marital.Status,Gender) %>% summarise(number = n()) %>% arrange(-number)

	Marital.Status	Gender	number
1	null	Male	3180
2	Single	Male	2314
3	Single	Female	1420
4	null	Female	1023
5	Married	Male	344
6	Married	Female	247
7	Widowed	Male	165
8	null	null	95
9	Widowed	Female	82
10	Divorced	Male	62
11	Divorced	Female	34
12	Single	null	30
13	Engaged	Female	19
14	Engaged	Male	19
15	null	Genderless	16
16	Separated	Male	16
17	Single	Genderless	15

ggplot(DN, aes(x = Marital.Status, fill = Gender)) + geom_bar(position = "fill")

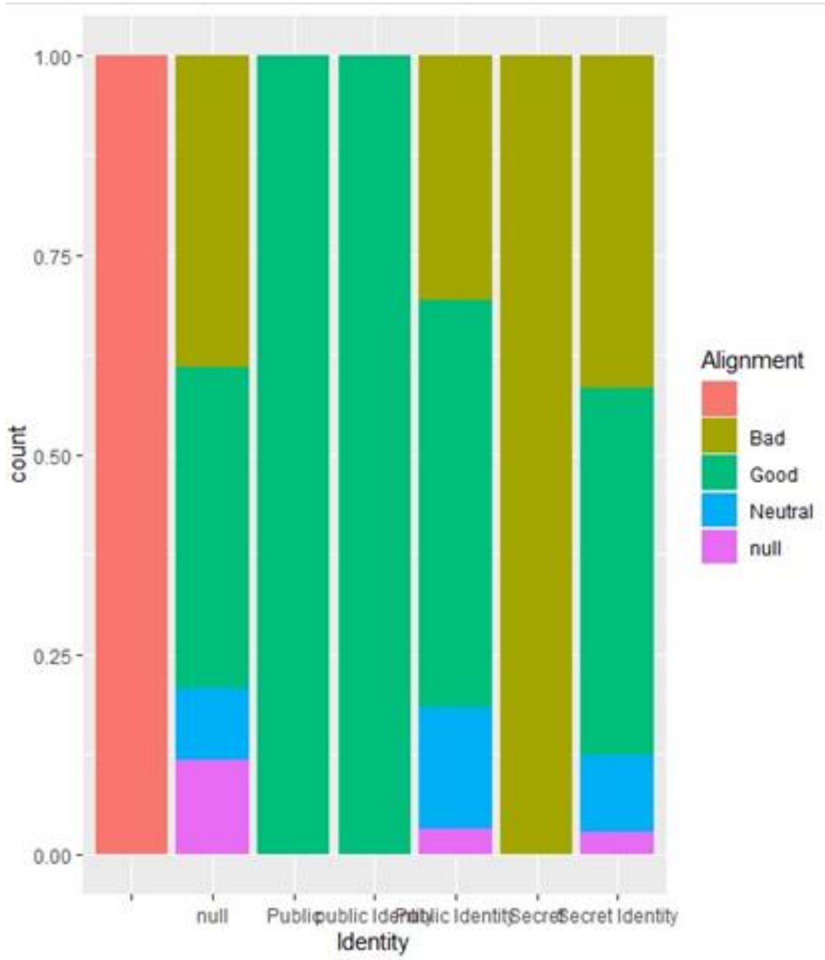
In this data analysis, the Characters percentage-wise represented each **marital status with gender segregation**. The colour emphasizes each gender-related to each status.



3.4 (Marital Status against Gender)

2) <u>Plot Identity against Alignment Variation of Each Character</u>

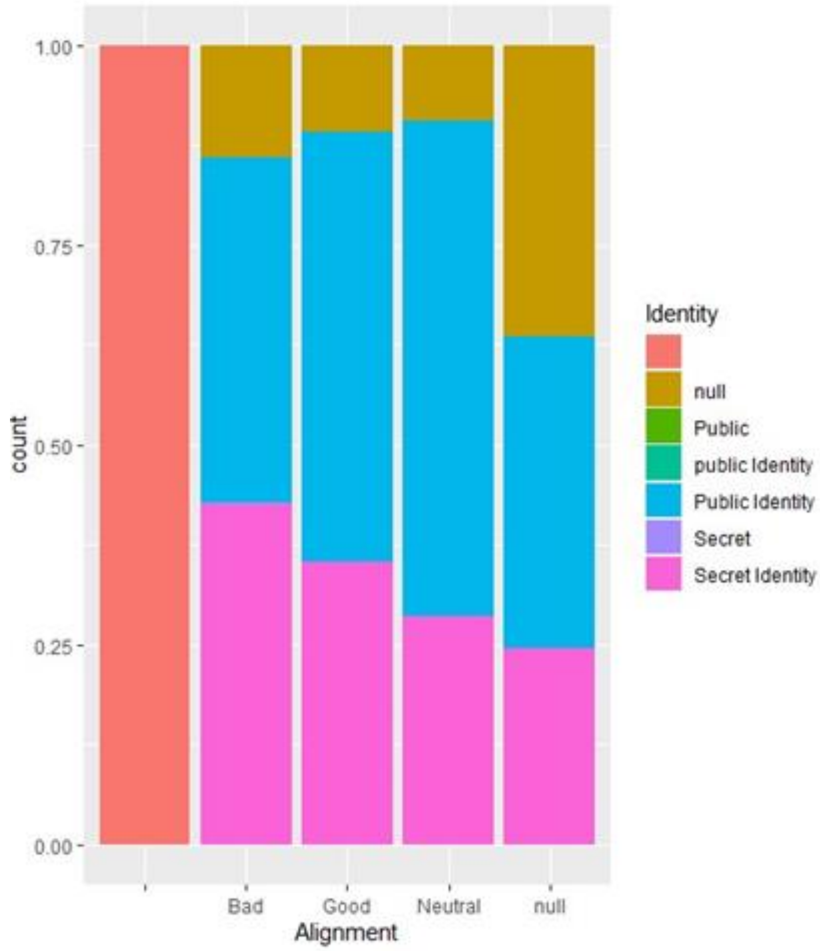
> ggplot(DC, aes(x = Identity, fill = Alignment)) + geom_bar(position = "fill")



3.5 (Proportion plot with Identity vs Count)

3) Plot Alignment against Identity Variation of Each Character

> ggplot(DC, aes(x = Alignment, fill = Identity)) + geom_bar(position = "fill")



3.6 - (Proportion plot with Alignment vs Count)

III) Frequency Plotting for Data Observation

```
DC<-read.csv("dc_2010_2020.csv")
listDC<-list()
summaryDC<-data.frame(matrix(vector(),ncol=5))
typeDC<-as.data.frame(unique(DC %>% filter(Identity=='Public Identity') %>%
select(Gender) %>% na.omit()))

colnames(summaryDC)<-typeDC
```

	Alignment	percentage_dc	number
1	Good	47.7	4352
2	Bad	35.7	3256
3	Neutral	12.3	1119
4	null	4.2	381
5		0.1	11

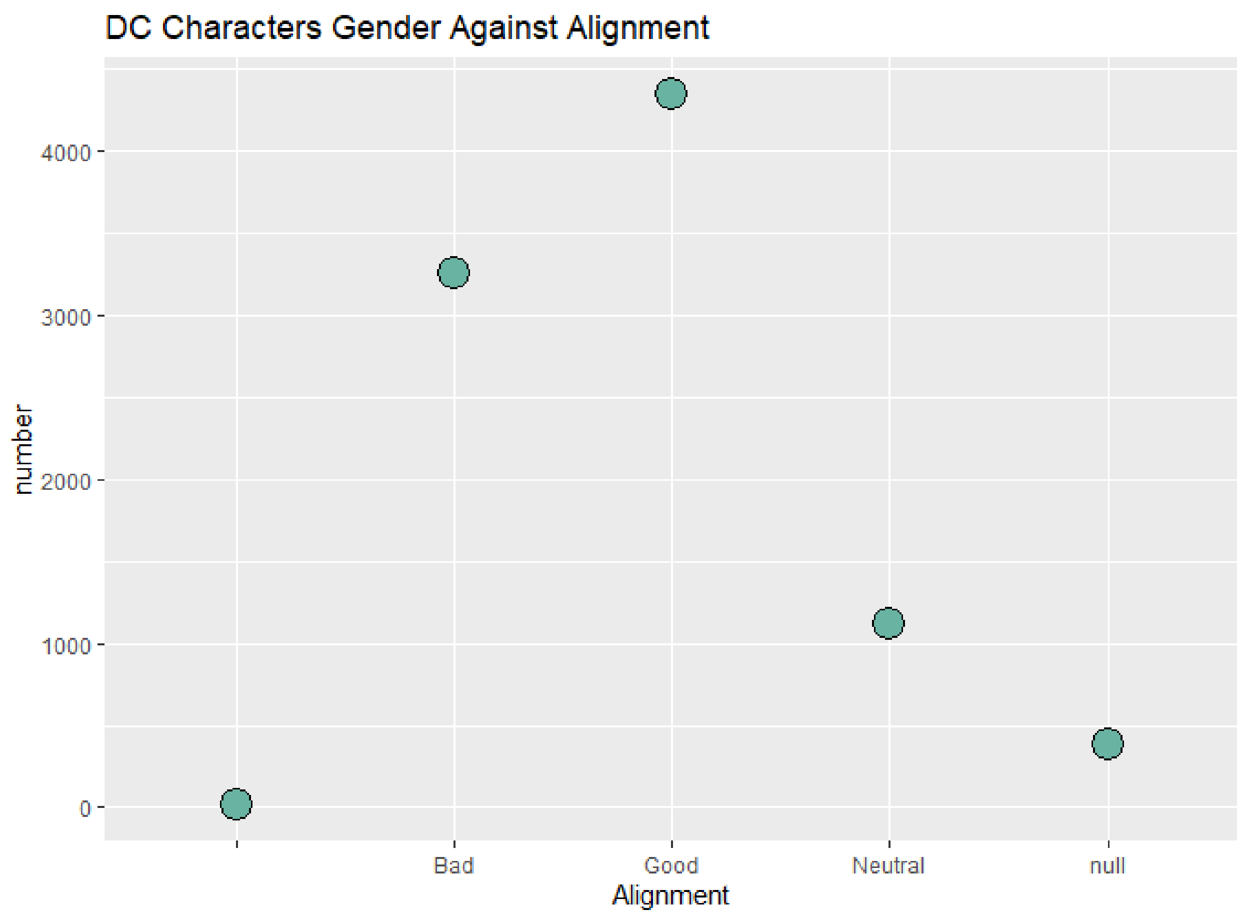
```
listDC<-as.data.frame(DC %>%
select(Alignment,Gender) %>% na.omit() %>%
group_by(Alignment) %>% summarise(number= n())
%>% arrange(-number) %>% mutate(countT=
sum(number)) %>%
mutate(percentage_dc=round(100*number/countT,1))
%>% select(Alignment,percentage_dc,number))
```

names(listDC)<-typeDC

Listing the characters according to Alignment basis and get the presentatge vise of Alignment from the whole population of DC Characters

```
library(hrbrthemes)

listDC %>% ggplot( aes(x=Alignment, y=number)) +
   geom_line( color="grey") +
   geom_point(shape=21, color="black", fill="#69b3a2", size=6) +
   ggtitle("DC Characters Gender Against Alignment")
```



3.7 - (Frequency plot for Characters Alignment)

In this plot , the data representation emphasizes the Alignment frequncy of the characters

```
>
> listDC<-list()
> summaryDC<-data.frame(matrix(vector(),ncol=5))
> typeDC<-as.data.frame(unique(DC %>% filter(Identity=='Public Identity') %>% select(Gender) %>% na.omit
()))
> colnames(summaryDC)<-typeDC
> listDC<-as.data.frame(DC %>%  select(Alignment,Gender) %>% na.omit() %>% group_by(Alignment) %>% summari
se(number= n()) %>% arrange(-number) %>% mutate(countT= sum(number)) %>% mutate(percentage_dc=round(100*nu
mber/countT,1)) %>% select(Alignment,percentage_dc,number))
> View(listDC)
> View(listDC)
> listDC %>% ggplot( aes(x=Alignment, y=number)) +
+     geom_line( color="grey") +
+     geom_point(shape=21, color="black", fill="#69b3a2", size=6) +
+     ggtitle("DC Characters Gender Against Alignment")
geom_path: Each group consists of only one observation. Do you need to adjust the group aesthetic?
```

4) Hypothesis Testing

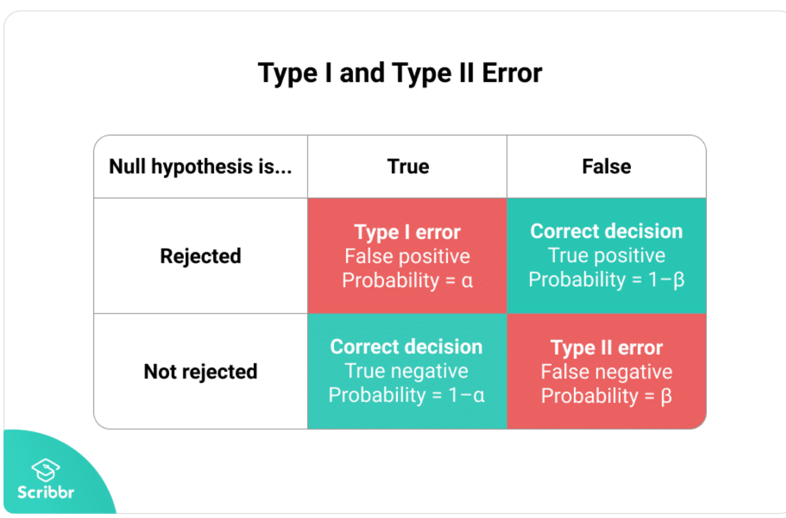
01. Character Analyzation with Identity & Gender

<u>In this hypothesis test</u>, we group characters with Identity and gender together and finding whether the data density of each grouped data accurate with Sample data

```
> xbar=DG$percentage[DG$Gender=="Male" & DG$Identity=="Public Identity"]
> DM=DC[1:1000,]
> DCAM<-DM %>% group_by(Gender,Identity) %>% summarise(number = n()) %>%
arrange(-number)
> DUM<-DCAM %>% group_by(Identity) %>% mutate(countT= sum(number)) %>%
group_by(Gender) %>% mutate(percentage=100*number/countT)
```

So the table that created Having columns with gender and identity. Mainly the identity is counted and from that we proportionated data for each gender and got overall percentages by gender wise.

In the hypothesis testing and find out whether we are doing a **type 1 or type 2 error** in the conclusion process.



Type I and Type II Error

Null hypothesis is...	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = 1−β
Not rejected	Correct decision True negative Probability = 1−α	Type II error False negative Probability = β

DC DM DUM listDC typeDC DCAI DU DG

Filter

	Gender	Identity	number	countT	percentage	[illegible]
1	Male	Public Identity	2973	4583	64.870172[illegible]8	64.87
2	Male	Secret Identity	2335	3344	69.82655502	[illegible]
3	Female	Public Identity	1540	4583	33.60244381	33.6
4	Female	Secret Identity	960	3344	28.70813397	28.71
5	Male	null	794	1177	67.45964316	67.46
6	Female	null	339	1177	28.80203908	28.8
7	null	Public Identity	52	4583	1.13462797	1.13
8	null	Secret Identity	38	3344	1.13636364	1.14
9	null	null	33	1177	2.80373832	2.8
10	Genderless	Public Identity	17	4583	0.37093607	0.37
11			11	11	100.00000000	100
12	Genderless	null	10	1177	0.84961767	0.85
13	Genderless	Secret Identity	4	3344	0.11961722	0.12
14	Transgender	Secret Identity	4	3344	0.11961722	0.12
15	Non-binary	Secret Identity	3	3344	0.08971292	0.09
16	null	Secret	2	2	100.00000000	100
17	Female	public identity	1	1	100.00000000	100
18	null	Public	1	1	100.00000000	100
19	Transgender	null	1	1177	0.08496177	0.08
20	Transgender	Public Identity	1	4583	0.02181977	0.02

4.1 (Percentage of Male DC characters who have a public identity is 64.87%.)

We got to know the percentage of Male DC characters who have a public identity is 64.87% in the whole data set so we say if we take 1000 samples from the dataset then in that sample there should be at least 64.87% of Male secret identity characters. At 5% significance level.

<u>Null Hypothesis</u>

If we take 1000 samples from the dataset then in that sample there should be at least 64.87% of Male secret identity characters.

<u>Alternative Hypothesis</u>

If we take 1000 samples from the dataset then in that sample there should be less than 64.87% of Male secret identity characters.

DC | listDC | typeDC | DCAI | DU | g1 | g2 | DCX | DM | DCNEW | DUM

Filter

	Gender	Identity	number	countT	percentage
1	Male	Secret Identity	326	442	73.7556561
2	Male	Public Identity	236	332	71.0843373
3	Male	null	159	223	71.3004484
4	Female	Secret Identity	107	442	24.2081448
5	Female	Public Identity	90	332	27.1084337
6	Female	null	57	223	25.5605381
7	null	Secret Identity	9	442	2.0361991
8	null	null	6	223	2.6905830
9	null	Public Identity	6	332	1.8072289
10	null	Secret	2	2	100.0000000
11			1	1	100.0000000
12	Genderless	null	1	223	0.4484305

> xbar=DG$percentage[DG$Gender=="Male" & DG$Identity=="Public Identity"]

> xbar

> mu0=DUM$percentage[DUM$Gender=="Male" & DUM$Identity=="Public Identity"]

> mu0

> sd(DG$number)

> sigma=sd(DG$number)

> z<-(xbar-mu0)/(sigma/sqrt(1000))

> p<-pnorm(z)

```
>  DM=DC[1:1000,]
> DM
   i..Year                                              Character                                                                 Character.href
1     2010                         Isabelle Rose Mahkent (New Earth)                      https://dc.fandom.com/wiki/Isabelle_Rose_Mahkent_(New_Earth)
2     2010                                      Ngo Sik (New Earth)                                    https://dc.fandom.com/wiki/Ngo_Sik_(New_Earth)
3     2010                                 Two-Ton Ted (New Earth)                                https://dc.fandom.com/wiki/Two-Ton_Ted_(New_Earth)
4     2010                Artemis of Bana-Mighdall (Superman/Batman)             https://dc.fandom.com/wiki/Artemis_of_Bana-Mighdall_(Superman/Batman)
5     2010                                Billy Batson (Earth-16)                                https://dc.fandom.com/wiki/Billy_Batson_(Earth-16)
6     2010                             Thomas Elliot (Hush Beyond)                             https://dc.fandom.com/wiki/Thomas_Elliot_(Hush_Beyond)
7     2010                                  Galahad II (New Earth)                                 https://dc.fandom.com/wiki/Galahad_II_(New_Earth)
8     2010                                     Medusa (Earth-508)                                     https://dc.fandom.com/wiki/Medusa_(Earth-508)
9     2010                    Mary Batson (The Brave and the Bold)                   https://dc.fandom.com/wiki/Mary_Batson_(The_Brave_and_the_Bold)
10    2010                       Darkseid (The Brave and the Bold)                      https://dc.fandom.com/wiki/Darkseid_(The_Brave_and_the_Bold)
11    2010                      Lionel Luthor (Smallville Earth-2)                     https://dc.fandom.com/wiki/Lionel_Luthor_(Smallville_Earth-2)
12    2010                               Mrs. Mercer (Smallville)                               https://dc.fandom.com/wiki/Mrs._Mercer_(Smallville)
13    2010                                     Gretel (Earth-508)                                     https://dc.fandom.com/wiki/Gretel_(Earth-508)
14    2010                                   Hunter II (New Earth)                                  https://dc.fandom.com/wiki/Hunter_II_(New_Earth)
15    2010           Tadwallader Jutefruce (The Brave and the Bold)          https://dc.fandom.com/wiki/Tadwallader_Jutefruce_(The_Brave_and_the_Bold)
16    2010                                     Bak Mei (New Earth)                                    https://dc.fandom.com/wiki/Bak_Mei_(New_Earth)
17    2010                              Rush Hour III (New Earth)                              https://dc.fandom.com/wiki/Rush_Hour_III_(New_Earth)
18    2010                     Chloroform (The Brave and the Bold)                    https://dc.fandom.com/wiki/Chloroform_(The_Brave_and_the_Bold)
19    2010                            Monsieur Mallah (Earth-508)                            https://dc.fandom.com/wiki/Monsieur_Mallah_(Earth-508)
20    2010                              Roderick Kane (New Earth)                              https://dc.fandom.com/wiki/Roderick_Kane_(New_Earth)
21    2010                        Gorilla Grodd (Joker's Playhouse)                    https://dc.fandom.com/wiki/Gorilla_Grodd_(Joker%27s_Playhouse)
22    2010                    Sweet Tooth (The Brave and the Bold)                   https://dc.fandom.com/wiki/Sweet_Tooth_(The_Brave_and_the_Bold)
23    2010                       Platinum (The Brave and the Bold)                      https://dc.fandom.com/wiki/Platinum_(The_Brave_and_the_Bold)
24    2010                   Walter Haley (The Brave and the Bold)                  https://dc.fandom.com/wiki/Walter_Haley_(The_Brave_and_the_Bold)
25    2010                  Herman Cramer (The Brave and the Bold)                 https://dc.fandom.com/wiki/Herman_Cramer_(The_Brave_and_the_Bold)
26    2010                     Bizarro Mister Miracle (New Earth)                     https://dc.fandom.com/wiki/Bizarro_Mister_Miracle_(New_Earth)
27    2010                                John Stewart (Earth-16)                                https://dc.fandom.com/wiki/John_Stewart_(Earth-16)
28    2010 Harley (Crisis on Two Earths: Crime Syndicate Earth) https://dc.fandom.com/wiki/Harley_(Crisis_on_Two_Earths:_Crime_Syndicate_Earth)
29    2010                  Arnold Wesker (The Brave and the Bold)                 https://dc.fandom.com/wiki/Arnold_Wesker_(The_Brave_and_the_Bold)
30    2010                                Lex Luthor (Tiny Titans)                               https://dc.fandom.com/wiki/Lex_Luthor_(Tiny_Titans)
31    2010                               Ming Dynasty (New Earth)                               https://dc.fandom.com/wiki/Ming_Dynasty_(New_Earth)
32    2010                       2-Face-2 (Batman in Bethlehem)                       https://dc.fandom.com/wiki/2-Face-2_(Batman_in_Bethlehem)
33    2010                James Gordon (Batman: Under the Red Hood)               https://dc.fandom.com/wiki/James_Gordon_(Batman:_Under_the_Red_Hood)
34    2010                               Darius Wayne (New Earth)                               https://dc.fandom.com/wiki/Darius_Wayne_(New_Earth)
```

```
> xbar=DG$percentage[DG$Gender=="Male" & DG$Identity=="Public Identity"]
> DM=DC[1:1000,]
> DCAM<-DM %>% group_by(Gender,Identity) %>% summarise(number = n()) %>% arrange(-number)
`summarise()` has grouped output by 'Gender'. You can override using the `.groups` argument.
> DUM<-DCAM %>% group_by(Identity) %>% mutate(countT= sum(number)) %>% group_by(Gender) %>% mutate(percentage=100*number/countT)
> xbar=DG$percentage[DG$Gender=="Male" & DG$Identity=="Public Identity"]
> xbar
[1] 64.87017
> mu0=DUM$percentage[DUM$Gender=="Male" & DUM$Identity=="Public Identity"]
> mu0
[1] 71.08434
> sd(DG$number)
[1] 863.8195
> sigma=sd(DG$number)
> z<-(xbar-mu0)/(sigma/sqrt(1000))
> z
[1] -0.2274887
> p<-pnorm(z)
> p
[1] 0.4100219
> |
```

Significance level = 5%

$$\alpha = 0.05$$

The P value of the above hypothesis testing is 0.41. **Which implies that the data is in valid range.**

p>α

Thus, If we take 1000 samples from the dataset then in that sample, the Percentage of British DC characters who have a secret identity is at least 64.87%. **At 5% significance level**.

Justification for Hypothesis

	Gender	Identity	number	countT	percentage
1	Male	Secret Identity	326	442	73.7556561
2	Male	Public Identity	236	332	71.0843373
3	Male	null	159	223	71.3004484
4	Female	Secret Identity	107	442	24.2081448
5	Female	Public Identity	90	332	27.1084337
6	Female	null	57	223	25.5605381
7	null	Secret Identity	9	442	2.0361991
8	null	null	6	223	2.6905830
9	null	Public Identity	6	332	1.8072289
10	null	Secret	2	2	100.0000000
11			1	1	100.0000000
12	Genderless	null	1	223	0.4484305

According to the above figure, we found out when we consider 1000 records, the percentage of Male DC characters who have a **secret identity is 71.300%**. 71.300%>64.87% this null hypothesis is valid. So the **alternative hypothesis** is rejected. This is a **Type 1 error** because first, we assume that the **null hypothesis** can be rejected but eventually this is a wrong assumption that we took

02. Character Analyzation with Identity & Citizenship

In this hypothesis test, we group characters with Identity and citizenship together and finding whether the data density of each grouped data accurate with Sample data

> DC<-read.csv("dc_2010_2020.csv")
> DCAT<-DC %>% group_by(Citizenship,Identity) %>% summarise(number = n()) %>% arrange(-number)

	Citizenship	Identity	number	countT	percentage
1	American	Public Identity	2406	4583	52.49836352
2	American	Secret Identity	1660	3344	49.64114833
3	null	Public Identity	1364	4583	29.76216452
4	null	Secret Identity	1157	3344	34.59928230
5	null	null	616	1177	52.33644860
6	American	null	402	1177	34.15463042
7	British	Public Identity	114	4583	2.48745363
8	British	Secret Identity	112	3344	3.34928230
9	Amazon	Public Identity	109	4583	2.37835479
10	Apokoliptian	Public Identity	74	4583	1.61466288
11	Atlantean	Public Identity	74	4583	1.61466288
12	Chinese	Secret Identity	41	3344	1.22607656
13	Genesisian	Public Identity	41	4583	0.89461052
14	Apokoliptian	null	27	1177	2.29396771
15	Apokoliptian	Secret Identity	27	3344	0.80741627
16	United Planets Citizen	Public Identity	27	4583	0.58913376
17	Russian	Public Identity	26	4583	0.56731399
18	Japanese	Secret Identity	25	3344	0.74760766
19	Atlantean	Secret Identity	24	3344	0.71770335
20	German	Public Identity	24	4583	0.52367445
21	Egyptian	Public Identity	23	4583	0.50185468
22	British	null	22	1177	1.86915888
23	Chinese	Public Identity	22	4583	0.48003491

4.2 (Percentage of British characters who have a Secret identity is 3.349%.)

> DT<-DCAT %>% group_by(Identity) %>% mutate(countT= sum(number)) %>% group_by(Citizenship) %>% mutate(percentage=100*number/countT)
> View(DT)

So, the table that created Having columns with Citizenship and identity. Mainly the Citizenship is counted and from that, we proportionated data for each Citizenship and got overall percentages by gender-wise.

- **(Percentage of British DC characters who have a secret identity is 3.349%.)**

We got to know the percentage of British DC characters who have a secret identity is 3.349% in the whole data set so we say if we take 1000 samples from the dataset then in that sample there should be at least 3.349% of British secret identity characters. At 5% significance level.

<u>Null Hypothesis</u>

If we take 1000 samples from the dataset then in that sample there should be at least 3.349% of British secret identity characters.

<u>Alternative Hypothesis</u>

If we take 1000 samples from the dataset then in that sample there should be less than 3.349% of British secret identity characters.

>xbar=DT$percentage[DT$Citizenship=="British" & DT$Identity=="Secret Identity"]
> DMT=DC[1:10,]
> DCAMT<-DMT %>% group_by(Citizenship,Identity) %>% summarise(number = n()) %>% arrange(-number)
`summarise()` has grouped output by 'Citizenship'. You can override using the `.groups` argument.
> View(DCAMT)
> DUMT<-DCAMT %>% group_by(Identity) %>% mutate(countT= sum(number)) %>% group_by(Citizenship) %>% mutate(percentage=100*number/countT)
> View(DUMT)

	Citizenship	Identity	number	countT	percentage
1	American	Secret Identity	190	442	42.9864253
2	American	Public Identity	184	332	55.4216867
3	null	Secret Identity	144	442	32.5791855
4	null	null	109	223	48.8789238
5	null	Public Identity	102	332	30.7228916
6	American	null	79	223	35.4260090
7	British	Secret Identity	55	442	12.4434389
8	Apokoliptian	null	8	223	3.5874439
9	Apokoliptian	Secret Identity	8	442	1.8099548
10	Chinese	Public Identity	7	332	2.1084337
11	Taiwanese	null	6	223	2.6905830
12	British	null	5	223	2.2421525
13	British	Public Identity	5	332	1.5060241
14	Atlantean	Secret Identity	4	442	0.9049774
15	Chinese	Secret Identity	4	442	0.9049774
16	French	Secret Identity	4	442	0.9049774
17	Hellion	Public Identity	4	332	1.2048193
18	Amazon	Secret Identity	3	442	0.6787330

4.3 (In a 1000 data Sample Identity Vs Citizenship Table)

> xbar
[1] 3.349282
> mu0=DUMT$percentage[DUMT$Citizenship=="British" & DUMT$Identity=="Secret Identity"]
> sd(DT$number)
[1] 226.4051
> sigma<-sd(DT$number)
> z<-(xbar-mu0)/(sigma/sqrt(1000))
> p<-pnorm(z)
> p
[1] 0.1020046
> z
[1] -1.270212

```
> DC<-read.csv("dc_2010_2020.csv")
> DCAT<-DC %>% group_by(Citizenship,Identity) %>% summarise(number = n()) %>% arrange(-number)
`summarise()` has grouped output by 'Citizenship'. You can override using the `.groups` argument.
> DT<-DCAT %>% group_by(Identity) %>% mutate(countT= sum(number)) %>% group_by(Citizenship) %>% mutate(percentage=100*numbe
r/countT)
>
> DMT=DC[1:1000,]
> xbar=DT$percentage[DT$Citizenship=="British" & DT$Identity=="Secret Identity"]
> DCAMT<-DMT %>% group_by(Citizenship,Identity) %>% summarise(number = n()) %>% arrange(-number)
`summarise()` has grouped output by 'Citizenship'. You can override using the `.groups` argument.
> DUMT<-DCAMT %>% group_by(Identity) %>% mutate(countT= sum(number)) %>% group_by(Citizenship) %>% mutate(percentage=100*num
ber/countT)
> xbar
[1] 3.349282
> mu0=DUMT$percentage[DUMT$Citizenship=="British" & DUMT$Identity=="Secret Identity"]
>
> sd(DT$number)
[1] 226.4051
> sigma<-sd(DT$number)
>
> z<-(xbar-mu0)/(sigma/sqrt(1000))
> p<-pnorm(z)
> p
[1] 0.1020046
> z
[1] -1.270212
```

Significance level = 5%

$$\alpha = 0.05$$

In this hypothesis testing, the P-Value is 0.102 which **implies that the data is in valid range**
$p > \alpha$

Thus If we take 1000 samples from the dataset then in that sample, the Percentage of British DC characters who have a secret identity is at least 3.349%. At 5% significance level.

Justification for Hypothesis

	Citizenship	Identity	number	countT	percentage
1	American	Secret Identity	190	442	42.9864253
2	American	Public Identity	184	332	55.4216867
3	null	Secret Identity	144	442	32.5791855
4	null	null	109	223	48.8789238
5	null	Public Identity	102	332	30.7228916
6	American	null	79	223	35.4260090
7	British	Secret Identity	55	442	12.4434389
8	Apokoliptian	null	8	223	3.5874439
9	Apokoliptian	Secret Identity	8	442	1.8099548
10	Chinese	Public Identity	7	332	2.1084337
11	Taiwanese	null	6	223	2.6905830
12	British	null	5	223	2.2421525
13	British	Public Identity	5	332	1.5060241
14	Atlantean	Secret Identity	4	442	0.9049774
15	Chinese	Secret Identity	4	442	0.9049774
16	French	Secret Identity	4	442	0.9049774
17	Hellion	Public Identity	4	332	1.2048193
18	Amazon	Secret Identity	3	442	0.6787330

- **According to the above figure we- found out when we consider 1000 records**, the percentage of British DC characters who have a secret identity is **12.443%**. 12.443%>3.349% this null hypothesis is valid. So the **alternative hypothesis** is rejected. This is a **Type 1 error** because first, we assume that the **null hypothesis** can be rejected but eventually this is a wrong assumption that we took

5) Plot the multivariate data

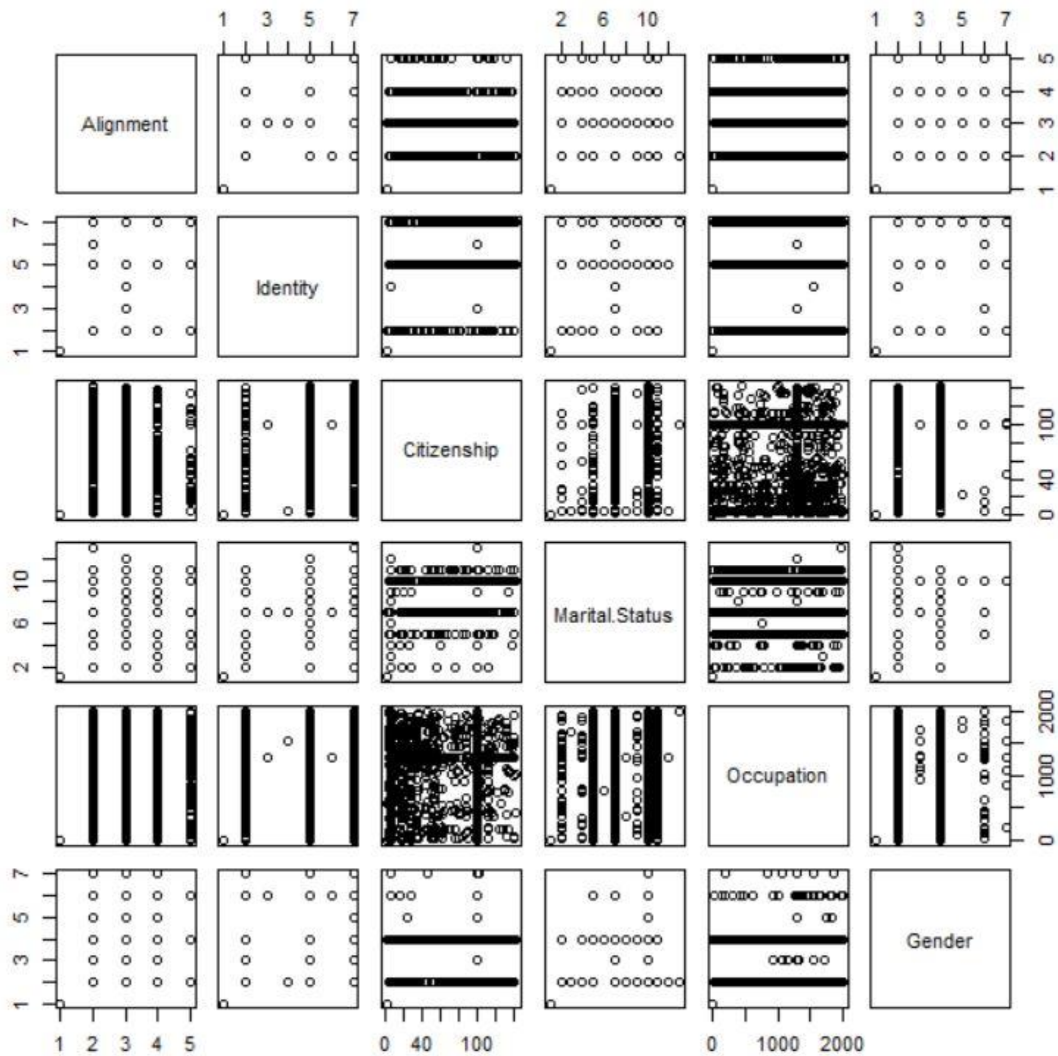
```
> dc <- read.csv("dc_2010_2020.csv",sep=",")
> head(dc)
```

```
> head(dc)
  i..Year                                  Character
1    2010          Isabelle Rose Mahkent (New Earth)
2    2010                        Ngo Sik (New Earth)
3    2010                    Two-Ton Ted (New Earth)
4    2010 Artemis of Bana-Mighdall (Superman/Batman)
5    2010                    Billy Batson (Earth-16)
6    2010                Thomas Elliot (Hush Beyond)
                                                 Character.href
1         https://dc.fandom.com/wiki/Isabelle_Rose_Mahkent_(New_Earth)
2                       https://dc.fandom.com/wiki/Ngo_Sik_(New_Earth)
3                   https://dc.fandom.com/wiki/Two-Ton_Ted_(New_Earth)
4 https://dc.fandom.com/wiki/Artemis_of_Bana-Mighdall_(Superman/Batman)
5                   https://dc.fandom.com/wiki/Billy_Batson_(Earth-16)
6               https://dc.fandom.com/wiki/Thomas_Elliot_(Hush_Beyond)
                Real.Name    Current.Alias Alignment        Identity
1   Isabelle Rose Mahkent Isabelle Mahkent   Neutral            null
2                 Ngo Sik          Go Seek       Bad            null
3                 Unknown      Two-Ton Ted       Bad Secret Identity
4 Artemis of Bana-Mighdall             null      Good            null
5  William "Billy" Batson           Shazam      Good Secret Identity
6           Thomas Elliot             Hush       Bad Public Identity
  Citizenship Marital.Status Occupation Gender  Hair  Eyes
1    American         Single       null Female  null  null
2  Vietnamese           null  Kidnapper   Male  null  null
3     British           null       null   Male  null  null
4      Amazon         Single       null Female   Red Green
5    American         Single Adventurer   Male Black  Blue
6    American         Single    Surgeon   Male   Red  Blue
                 Universe                          First.Appearance
1                    null       JSA All-Stars  #11\n(December, 2010)
2               New Earth              Azrael Vol 2  #7\n(June, 2010)
3               New Earth    Knight and Squire  #1\n(December, 2010)
4 Superman/Batman (Reality)                                     null
5                Earth-16                                       null
6             Hush Beyond Batman Beyond Vol 3  #2\n(September, 2010)
                  Appearance.of.Death
1                                null
2           Azrael Vol 2  #7\n(June, 2010)
3                                null
4                                null
5                                null
6 Batman Beyond Vol 3  #5\n(December, 2010)
> 
```

We have used above code segments to read multivariate data.

```
plot(dc[6:11])
```

Using this 'plot(dc[6:11]) ' command we have plotted the multivariate data between column 6 and column 11. Here we have plotted graphs for the following data. **(Alignment, Identity, Citizenship, Material Status, Occupation and Gender)**



4.1 (Multivariate plot Respective to Whole DC-Character Dataset)

6)Relationship between Variables

I) Cor-relation

> DCAU<-DC %>% group_by(Gender,Identity) %>% summarise(number = n()) %>% arrange(-number)

> DG<-DCAU %>% group_by(Identity) %>% mutate(countT= sum(number)) %>% group_by(Gender) %>% mutate(percentage=100*number/countT)

> DG$LABEL <-paste0(round(DG$percentage,2))

	Gender	Identity	number	countT	percentage	LABEL
1	Male	Public Identity	2973	4583	64.87017238	64.87
2	Male	Secret Identity	2335	3344	69.82655502	69.83
3	Female	Public Identity	1540	4583	33.60244381	33.6
4	Female	Secret Identity	960	3344	28.70813397	28.71
5	Male	null	794	1177	67.45964316	67.46
6	Female	null	339	1177	28.80203908	28.8
7	null	Public Identity	52	4583	1.13462797	1.13
8	null	Secret Identity	38	3344	1.13636364	1.14
9	null	null	33	1177	2.80373832	2.8
10	Genderless	Public Identity	17	4583	0.37093607	0.37

6.1 - (View of DG that represent Gender with Identity)

To find the Correlations we wanted to create independent variable and depended variable with common grouping in the dataset. So, our intentions are to grouping DC Characters with Identity combine **Gender and Alignment**.

	Alignment	Identity	number	countT	percentage	LABEL
1	Good	Public Identity	2336	4583	50.970980	50.97
2	Good	Secret Identity	1542	3344	46.112440	46.11
3	Bad	Public Identity	1405	4583	30.656775	30.66
4	Bad	Secret Identity	1389	3344	41.537081	41.54
5	Neutral	Public Identity	694	4583	15.142919	15.14
6	Good	null	472	1177	40.101954	40.1
7	Bad	null	460	1177	39.082413	39.08
8	Neutral	Secret Identity	319	3344	9.539474	9.54
9	null	Public Identity	148	4583	3.229326	3.23
10	null	null	139	1177	11.809686	11.81
11	Neutral	null	106	1177	9.005947	9.01
12	null	Secret Identity	94	3344	2.811005	2.81
13			11	11	100.000000	100
14	Bad	Secret	2	2	100.000000	100
15	Good	Public	1	1	100.000000	100
16	Good	public Identity	1	1	100.000000	100

6.2 (View of DU that represent Alignment with Identity)

```
> JoinGA=merge(x=DG,y=DU,by="Identity",all=TRUE)
> head(JoinGA)
  Identity Gender number.x countT.x percentage.x LABEL.x Alignment number.y countT.y percentage.y LABEL.y
1                      11       11   100.000000     100                 11       11   100.000000     100
2     null   Male     794     1177    67.459643   67.46      Good      472     1177    40.101954    40.1
3     null   Male     794     1177    67.459643   67.46       Bad      460     1177    39.082413   39.08
4     null   Male     794     1177    67.459643   67.46      null      139     1177    11.809686   11.81
5     null   Male     794     1177    67.459643   67.46   Neutral      106     1177     9.005947    9.01
6     null   null      33     1177     2.803738     2.8      Good      472     1177    40.101954    40.1
> plot(JoinGA[2:6])
> plot(JoinGA[7:11])
> |
```

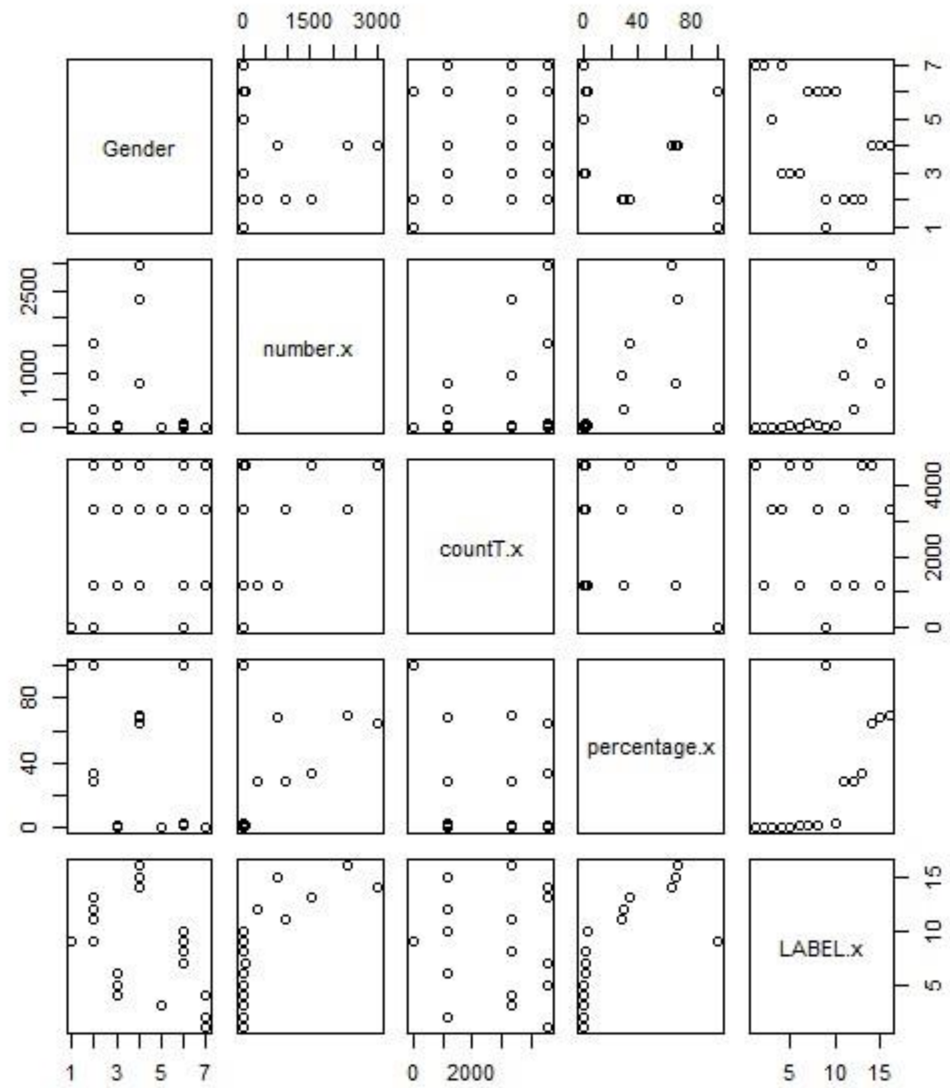
	Identity	Gender	number.x	countT.x	percentage.x	LABEL.x	Alignment	number.y	countT.y	percentage.y	LABEL.y
1			11	11	100.00000000	100		11	11	100.000000	100
2	null	Male	794	1177	67.45964316	67.46	Good	472	1177	40.101954	40.1
3	null	Male	794	1177	67.45964316	67.46	Bad	460	1177	39.082413	39.08
4	null	Male	794	1177	67.45964316	67.46	null	139	1177	11.809686	11.81
5	null	Male	794	1177	67.45964316	67.46	Neutral	106	1177	9.005947	9.01
6	null	null	33	1177	2.80373832	2.8	Good	472	1177	40.101954	40.1
7	null	null	33	1177	2.80373832	2.8	Bad	460	1177	39.082413	39.08
8	null	null	33	1177	2.80373832	2.8	null	139	1177	11.809686	11.81
9	null	null	33	1177	2.80373832	2.8	Neutral	106	1177	9.005947	9.01
10	null	Female	339	1177	28.80203908	28.8	Good	472	1177	40.101954	40.1
11	null	Female	339	1177	28.80203908	28.8	Bad	460	1177	39.082413	39.08
12	null	Female	339	1177	28.80203908	28.8	null	139	1177	11.809686	11.81
13	null	Female	339	1177	28.80203908	28.8	Neutral	106	1177	9.005947	9.01
14	null	Transgender	1	1177	0.08496177	0.08	Good	472	1177	40.101954	40.1
15	null	Transgender	1	1177	0.08496177	0.08	Bad	460	1177	39.082413	39.08
16	null	Transgender	1	1177	0.08496177	0.08	null	139	1177	11.809686	11.81
17	null	Transgender	1	1177	0.08496177	0.08	Neutral	106	1177	9.005947	9.01
18	null	Genderless	10	1177	0.84961767	0.85	Good	472	1177	40.101954	40.1
19	null	Genderless	10	1177	0.84961767	0.85	Bad	460	1177	39.082413	39.08
20	null	Genderless	10	1177	0.84961767	0.85	null	139	1177	11.809686	11.81
21	null	Genderless	10	1177	0.84961767	0.85	Neutral	106	1177	9.005947	9.01
22	Public	null	1	1	100.00000000	100	Good	1	1	100.000000	100
23	public Identity	Female	1	1	100.00000000	100	Good	1	1	100.000000	100
24	Public Identity	Male	2973	4583	64.87017238	64.87	Good	2336	4583	50.970980	50.97
25	Public Identity	Male	2973	4583	64.87017238	64.87	Bad	1405	4583	30.656775	30.66
26	Public Identity	Male	2973	4583	64.87017238	64.87	Neutral	694	4583	15.142919	15.14
27	Public Identity	Male	2973	4583	64.87017238	64.87	null	148	4583	3.229326	3.23
28	Public Identity	Genderless	17	4583	0.37093607	0.37	Good	2336	4583	50.970980	50.97
29	Public Identity	Genderless	17	4583	0.37093607	0.37	Bad	1405	4583	30.656775	30.66
30	Public Identity	Genderless	17	4583	0.37093607	0.37	Neutral	694	4583	15.142919	15.14
31	Public Identity	Genderless	17	4583	0.37093607	0.37	null	148	4583	3.229326	3.23
32	Public Identity	Female	1540	4583	33.60244381	33.6	Good	2336	4583	50.970980	50.97
33	Public Identity	Female	1540	4583	33.60244381	33.6	Bad	1405	4583	30.656775	30.66

C:/Users/User/Desktop/R/

32	Public Identity	Female	1540	4583	33.60244381	33.6	Good	2336	4583	50.970980	50.97
33	Public Identity	Female	1540	4583	33.60244381	33.6	Bad	1405	4583	30.656775	30.66
34	Public Identity	Female	1540	4583	33.60244381	33.6	Neutral	694	4583	15.142919	15.14
35	Public Identity	Female	1540	4583	33.60244381	33.6	null	148	4583	3.229326	3.23
36	Public Identity	Transgender	1	4583	0.02181977	0.02	Good	2336	4583	50.970980	50.97
37	Public Identity	Transgender	1	4583	0.02181977	0.02	Bad	1405	4583	30.656775	30.66
38	Public Identity	Transgender	1	4583	0.02181977	0.02	Neutral	694	4583	15.142919	15.14
39	Public Identity	Transgender	1	4583	0.02181977	0.02	null	148	4583	3.229326	3.23
40	Public Identity	null	52	4583	1.13462797	1.13	Good	2336	4583	50.970980	50.97
41	Public Identity	null	52	4583	1.13462797	1.13	Bad	1405	4583	30.656775	30.66
42	Public Identity	null	52	4583	1.13462797	1.13	Neutral	694	4583	15.142919	15.14
43	Public Identity	null	52	4583	1.13462797	1.13	null	148	4583	3.229326	3.23
44	Secret	null	2	2	100.00000000	100	Bad	2	2	100.000000	100
45	Secret Identity	Genderless	4	3344	0.11961722	0.12	Good	1542	3344	46.112440	46.11
46	Secret Identity	Genderless	4	3344	0.11961722	0.12	Bad	1389	3344	41.537081	41.54
47	Secret Identity	Genderless	4	3344	0.11961722	0.12	Neutral	319	3344	9.539474	9.54
48	Secret Identity	Genderless	4	3344	0.11961722	0.12	null	94	3344	2.811005	2.81
49	Secret Identity	Male	2335	3344	69.82655502	69.83	Good	1542	3344	46.112440	46.11
50	Secret Identity	Male	2335	3344	69.82655502	69.83	Bad	1389	3344	41.537081	41.54
51	Secret Identity	Male	2335	3344	69.82655502	69.83	Neutral	319	3344	9.539474	9.54
52	Secret Identity	Male	2335	3344	69.82655502	69.83	null	94	3344	2.811005	2.81
53	Secret Identity	Female	960	3344	28.70813397	28.71	Good	1542	3344	46.112440	46.11
54	Secret Identity	Female	960	3344	28.70813397	28.71	Bad	1389	3344	41.537081	41.54
55	Secret Identity	Female	960	3344	28.70813397	28.71	Neutral	319	3344	9.539474	9.54
56	Secret Identity	Female	960	3344	28.70813397	28.71	null	94	3344	2.811005	2.81
57	Secret Identity	Transgender	4	3344	0.11961722	0.12	Good	1542	3344	46.112440	46.11
58	Secret Identity	Transgender	4	3344	0.11961722	0.12	Bad	1389	3344	41.537081	41.54
59	Secret Identity	Transgender	4	3344	0.11961722	0.12	Neutral	319	3344	9.539474	9.54
60	Secret Identity	Transgender	4	3344	0.11961722	0.12	null	94	3344	2.811005	2.81
61	Secret Identity	Non-binary	3	3344	0.08971292	0.09	Good	1542	3344	46.112440	46.11
62	Secret Identity	Non-binary	3	3344	0.08971292	0.09	Bad	1389	3344	41.537081	41.54
63	Secret Identity	Non-binary	3	3344	0.08971292	0.09	Neutral	319	3344	9.539474	9.54
64	Secret Identity	Non-binary	3	3344	0.08971292	0.09	null	94	3344	2.811005	2.81
65	Secret Identity	null	38	3344	1.13636364	1.14	Good	1542	3344	46.112440	46.11
66	Secret Identity	null	38	3344	1.13636364	1.14	Bad	1389	3344	41.537081	41.54
67	Secret Identity	null	38	3344	1.13636364	1.14	Neutral	319	3344	9.539474	9.54
68	Secret Identity	null	38	3344	1.13636364	1.14	null	94	3344	2.811005	2.81

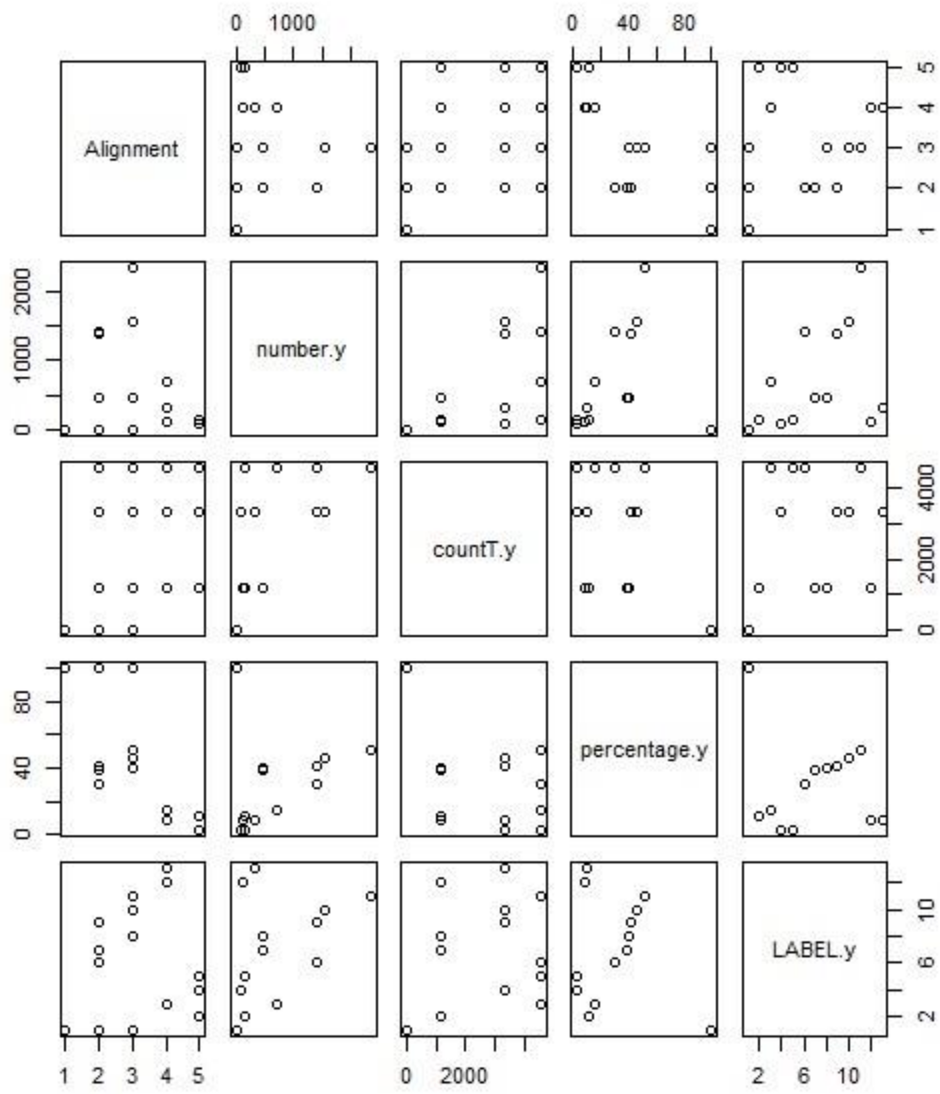
6.3 (View of Joined Table)

- Using this 'plot (JoinGA[2:6]) ' command we have plotted the multivariate data between column 2 and column 6. Here we have plotted graphs for the following data. (Gender, number.x, CountT.x, percentage.x, and Label.x)



6.4 (Multivariate plot from Joined table)

- Using this ‘ plot(JoinGA[7:11]) ‘ command we have plotted the multivariate data between column 7 and column 11. Here we have plotted graphs for the following data. (Alignment, number.y, CountT.y, percentage.y, and Label.y)



6.5 (Multivariate plot from Joined table range 7:11)

> Gender=JoinGA$number.x

<u>(Get gender counts from join table for correlation data analysis)</u>

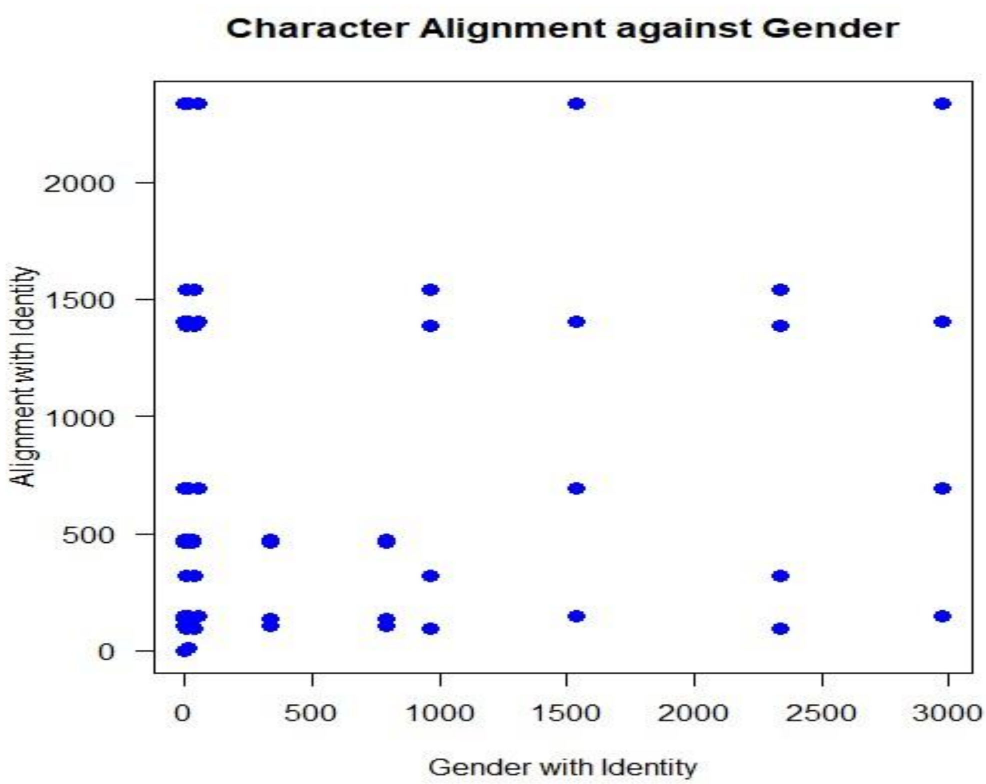
```
> Gender=JoinGA$number.x
> Gender
 [1]   11  794  794  794  794   33   33   33   33  339  339  339  339    1    1    1    1   10   10   10   10    1    1 2973 2973 2973 2973   17   17   17   17
[32] 1540 1540 1540 1540    1    1    1    1   52   52   52   52    2    4    4    4    4 2335 2335 2335 2335  960  960  960  960    4    4    4    4    3    3
[63]    3    3   38   38   38   38
```

> Alignment=JoinGA$number.y

(Get alignment counts from join table for correlation analysis)

```
> Alignment=JoinGA$number.y
> Alignment
 [1]   11  472  460  139  106  472  460  139  106  472  460  139  106  472  460  139  106  472  460  139  106    1    1 2336 1405  694  148 2336 1405  694  148
[32] 2336 1405  694  148 2336 1405  694  148 2336 1405  694  148    2 1542 1389  319   94 1542 1389  319   94 1542 1389  319   94 1542 1389  319   94 1542 1389
[63]  319   94 1542 1389  319   94
```

> p= plot(Gender,Alignment,xlab="Gender with Identity",ylab="Alignment with Identity",main="Character Alignment against Gender",pch=16,cex=1.3,col="blue",las=1)



6.6 (Plot of character Alignment against Gender)

```
> cor(Gender,Alignment,method="pearson")
[1] 0.1742544
> cor.test(Gender,Alignment,method="pearson")

        Pearson's product-moment correlation

data:  Gender and Alignment
t = 1.4376, df = 66, p-value = 0.1553
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.06695237  0.39621795
sample estimates:
      cor
0.1742544
```

According to the cor.test we are getting 0.1742 which is very low relationship because R square is 0.030 which means 3.0% of alignment data can be represented from given Gender.

II)Regression Line

Create the regression line,with loaded coefficients.

> LSRL<-lm(Alignment~Gender)

> p_LSRL=plot(Gender,Alignment,xlab="Gender with Identity",ylab="Alignment with Identity",main="Least Square Regression line Plot",pch=16,cex=1.3,col="black")

> abline(coefficients(LSRL), lwd=2, lty=2,col="red")

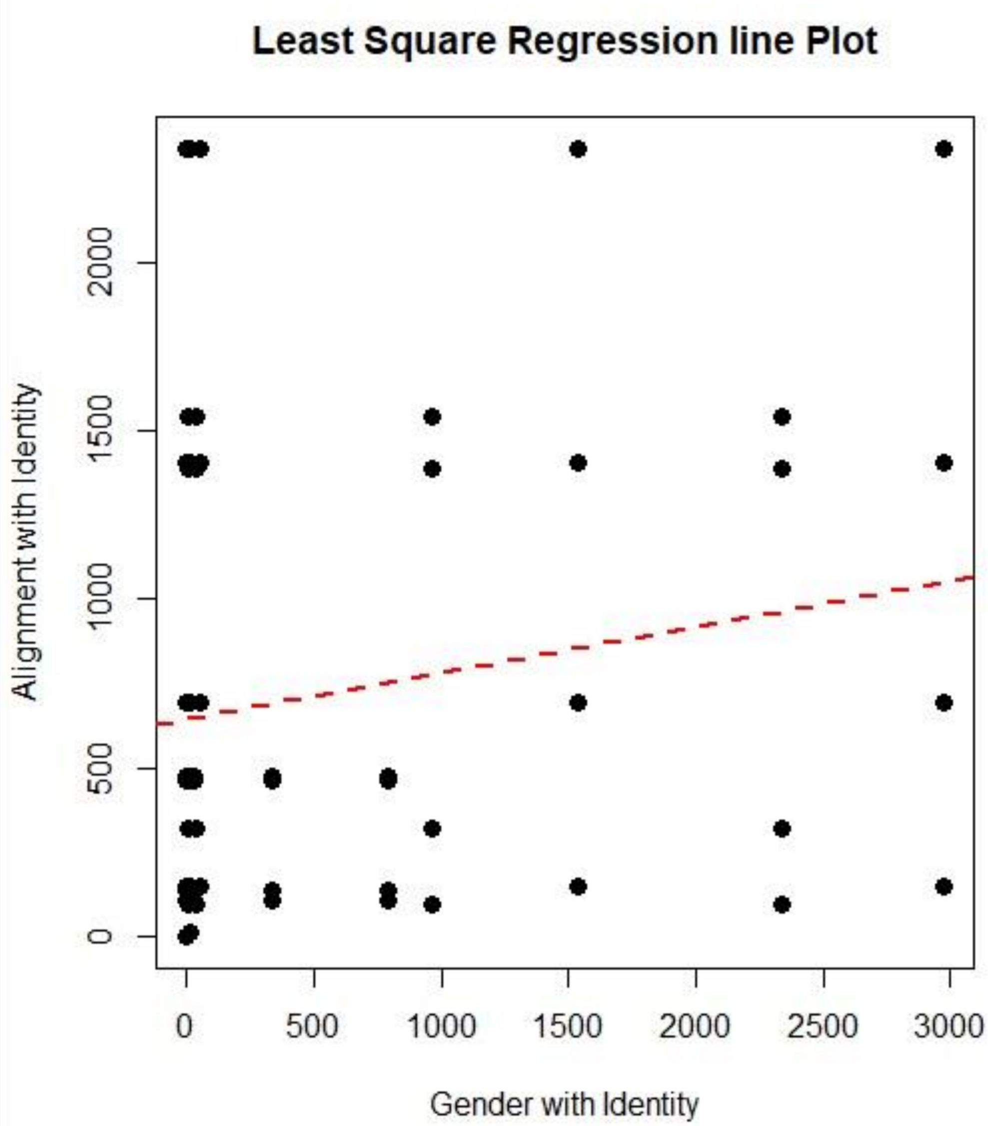
```
> LSRL<-lm(Alignment~Gender)
> p_LSRL=plot(Gender,Alignment,xlab="Gender with Identity",ylab="Alignment with Identity",main="Least Square Regression line Plot",pch=16,cex=1.3,col="black"
> abline(coefficients(LSRL), lwd=2, lty=2,col="red")
>
> LSRL

Call:
lm(formula = Alignment ~ Gender)

Coefficients:
(Intercept)       Gender
   645.3014       0.1372

> |
```

In this data representation we creating the regression line according to the Gender and Alignment. **Gender= Independent Variable**
Alignment = Dependent Variable



6.7 (Plot of least square regression line)

III) <u>Residual Plot</u>

- **Create the model with Alignment and Gender. Which takes x=Gender and y=Alignment.**

> Alignment.lm=lm(Alignment~Gender)

Get residuals with-above model.

> Alignment.res=resid(Alignment.lm)

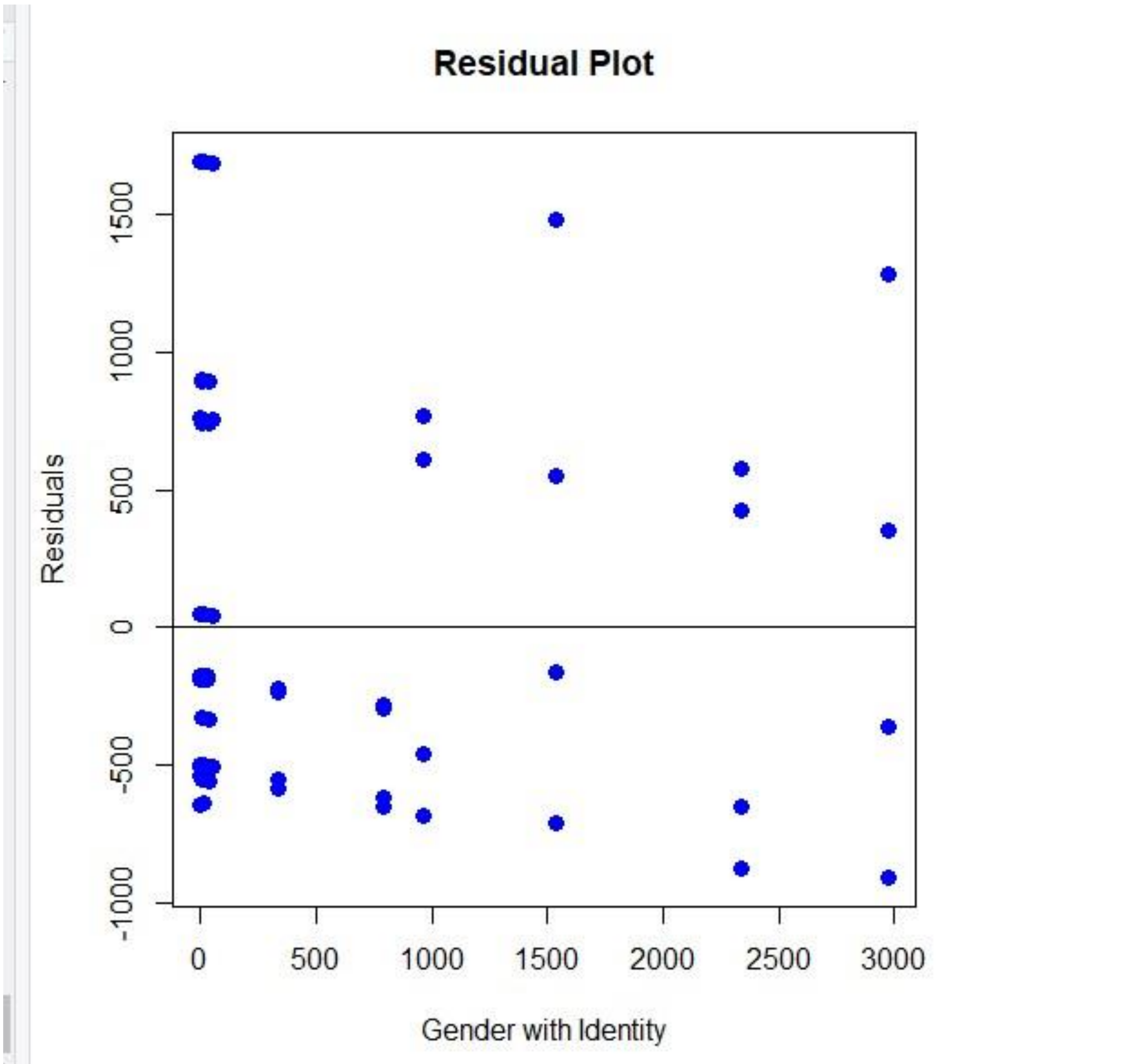
Make the residual plot.

> p_resid=plot(Gender,Alignment.res,xlab="Gender with Identity",ylab="Residuals",main="Residual Plot", pch=16,cex=1.3,col="blue")

Draw the (0,0) line.

> abline(0,0)

```
> Alignment.lm=lm(Alignment~Gender)
> Alignment.res=resid(Alignment.lm)
> p_resid=plot(Gender,Alignment.res,xlab="Gender with Identity",ylab="Residuals",main="Residual Plot",pch=16,cex=1.3,col="blue")
> abline(0,0)
> |
```



6.8 (Residual plot for gender with Identity)

By looking at the residual plot there **is no independent pattern**. So, from this we can get a conclusion like there is a **non-constant** variance between those two variables.

7)<u>Hierarchical Clustering</u>

JoinGA Table for Cluster Identity, Gender and Alignment of Dc Comic Characters.

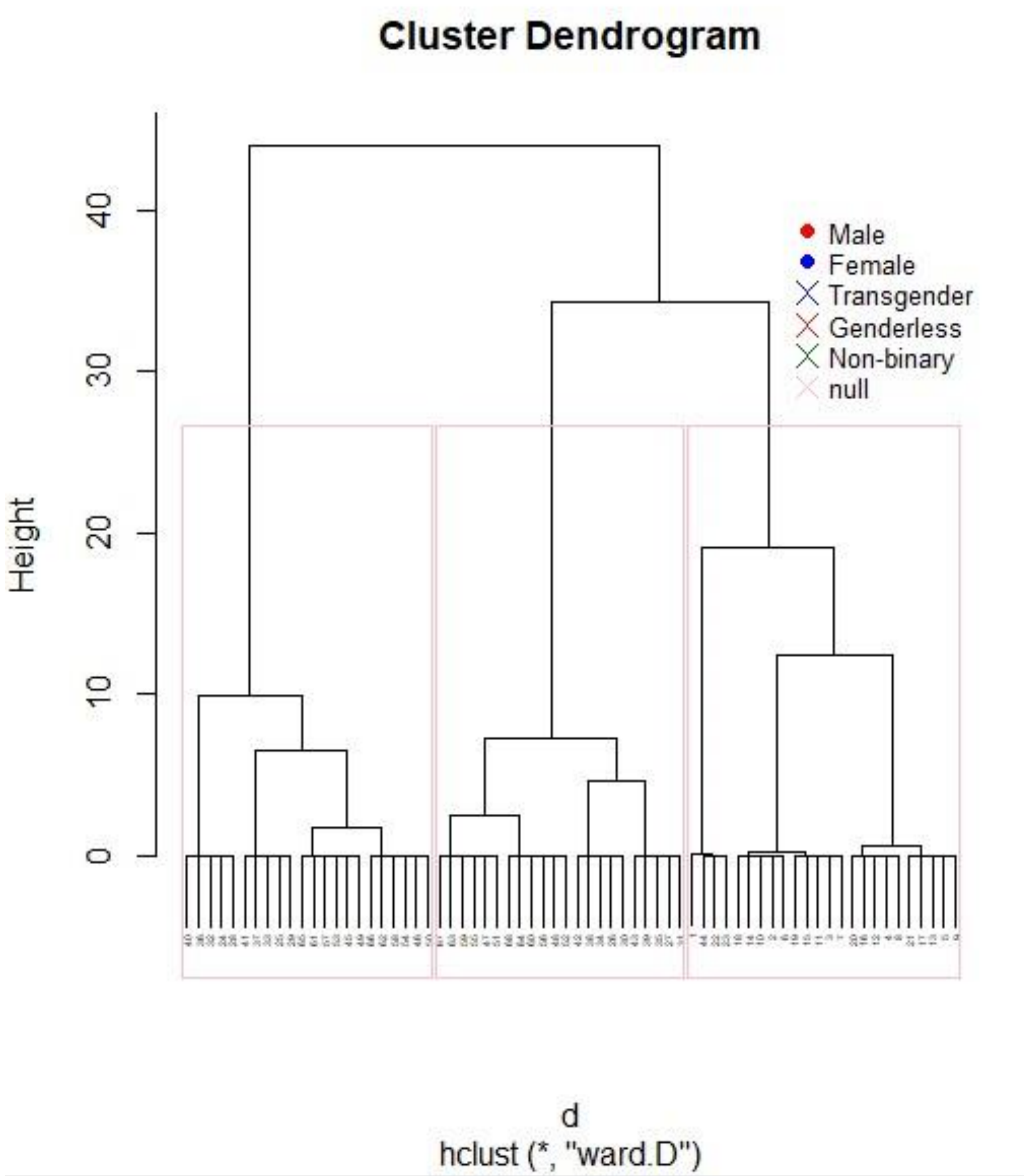
	Identity	Gender	number.x	countT.x	percentage.x	LABEL.x	Alignment	number.y	countT.y	percentage.y	LABEL.y
1			11	11	100.00000000	100		11	11	100.000000	100
2	null	Male	794	1177	67.45964316	67.46	Good	472	1177	40.101954	40.1
3	null	Male	794	1177	67.45964316	67.46	Bad	460	1177	39.082413	39.08
4	null	Male	794	1177	67.45964316	67.46	null	139	1177	11.809686	11.81
5	null	Male	794	1177	67.45964316	67.46	Neutral	106	1177	9.005947	9.01
6	null	null	33	1177	2.80373832	2.8	Good	472	1177	40.101954	40.1
7	null	null	33	1177	2.80373832	2.8	Bad	460	1177	39.082413	39.08
8	null	null	33	1177	2.80373832	2.8	null	139	1177	11.809686	11.81
9	null	null	33	1177	2.80373832	2.8	Neutral	106	1177	9.005947	9.01
10	null	Female	339	1177	28.80203908	28.8	Good	472	1177	40.101954	40.1
11	null	Female	339	1177	28.80203908	28.8	Bad	460	1177	39.082413	39.08
12	null	Female	339	1177	28.80203908	28.8	null	139	1177	11.809686	11.81
13	null	Female	339	1177	28.80203908	28.8	Neutral	106	1177	9.005947	9.01
14	null	Transgender	1	1177	0.08496177	0.08	Good	472	1177	40.101954	40.1
15	null	Transgender	1	1177	0.08496177	0.08	Bad	460	1177	39.082413	39.08
16	null	Transgender	1	1177	0.08496177	0.08	null	139	1177	11.809686	11.81
17	null	Transgender	1	1177	0.08496177	0.08	Neutral	106	1177	9.005947	9.01
18	null	Genderless	10	1177	0.84961767	0.85	Good	472	1177	40.101954	40.1
19	null	Genderless	10	1177	0.84961767	0.85	Bad	460	1177	39.082413	39.08
20	null	Genderless	10	1177	0.84961767	0.85	null	139	1177	11.809686	11.81
21	null	Genderless	10	1177	0.84961767	0.85	Neutral	106	1177	9.005947	9.01
22	Public	null	1	1	100.00000000	100	Good	1	1	100.000000	100
23	public Identity	Female	1	1	100.00000000	100	Good	1	1	100.000000	100
24	Public Identity	Male	2973	4583	64.87017238	64.87	Good	2336	4583	50.970980	50.97
25	Public Identity	Male	2973	4583	64.87017238	64.87	Bad	1405	4583	30.656775	30.66
26	Public Identity	Male	2973	4583	64.87017238	64.87	Neutral	694	4583	15.142919	15.14
27	Public Identity	Male	2973	4583	64.87017238	64.87	null	148	4583	3.229326	3.23
28	Public Identity	Genderless	17	4583	0.37093607	0.37	Good	2336	4583	50.970980	50.97
29	Public Identity	Genderless	17	4583	0.37093607	0.37	Bad	1405	4583	30.656775	30.66
30	Public Identity	Genderless	17	4583	0.37093607	0.37	Neutral	694	4583	15.142919	15.14
31	Public Identity	Genderless	17	4583	0.37093607	0.37	null	148	4583	3.229326	3.23
32	Public Identity	Female	1540	4583	33.60244381	33.6	Good	2336	4583	50.970980	50.97
33	Public Identity	Female	1540	4583	33.60244381	33.6	Bad	1405	4583	30.656775	30.66

Hierarchical clustering method is appropriate for the above data manipulation to percentagewise find out the Each Characters Identity, Gender and Alignment group together. From the created denodgram Characters **Alignment with identity will be segregated** from whole population and do the same ; **Gender respective with Identity**.

```
C:/Users/User/Desktop/R/
32 Public Identity      Female  1540  4583  33.60244381  33.6     Good  2336  4583  50.970980  50.97
33 Public Identity      Female  1540  4583  33.60244381  33.6      Bad  1405  4583  30.656775  30.66
34 Public Identity      Female  1540  4583  33.60244381  33.6  Neutral   694  4583  15.142919  15.14
35 Public Identity      Female  1540  4583  33.60244381  33.6     null   148  4583   3.229326   3.23
36 Public Identity Transgender     1  4583   0.02181977  0.02     Good  2336  4583  50.970980  50.97
37 Public Identity Transgender     1  4583   0.02181977  0.02      Bad  1405  4583  30.656775  30.66
38 Public Identity Transgender     1  4583   0.02181977  0.02  Neutral   694  4583  15.142919  15.14
39 Public Identity Transgender     1  4583   0.02181977  0.02     null   148  4583   3.229326   3.23
40 Public Identity        null    52  4583   1.13462797  1.13     Good  2336  4583  50.970980  50.97
41 Public Identity        null    52  4583   1.13462797  1.13      Bad  1405  4583  30.656775  30.66
42 Public Identity        null    52  4583   1.13462797  1.13  Neutral   694  4583  15.142919  15.14
43 Public Identity        null    52  4583   1.13462797  1.13     null   148  4583   3.229326   3.23
44          Secret        null     2     2 100.00000000   100      Bad     2     2 100.000000    100
45 Secret Identity  Genderless     4  3344   0.11961722  0.12     Good  1542  3344  46.112440  46.11
46 Secret Identity  Genderless     4  3344   0.11961722  0.12      Bad  1389  3344  41.537081  41.54
47 Secret Identity  Genderless     4  3344   0.11961722  0.12  Neutral   319  3344   9.539474   9.54
48 Secret Identity  Genderless     4  3344   0.11961722  0.12     null    94  3344   2.811005   2.81
49 Secret Identity        Male  2335  3344  69.82655502 69.83     Good  1542  3344  46.112440  46.11
50 Secret Identity        Male  2335  3344  69.82655502 69.83      Bad  1389  3344  41.537081  41.54
51 Secret Identity        Male  2335  3344  69.82655502 69.83  Neutral   319  3344   9.539474   9.54
52 Secret Identity        Male  2335  3344  69.82655502 69.83     null    94  3344   2.811005   2.81
53 Secret Identity      Female   960  3344  28.70813397 28.71     Good  1542  3344  46.112440  46.11
54 Secret Identity      Female   960  3344  28.70813397 28.71      Bad  1389  3344  41.537081  41.54
55 Secret Identity      Female   960  3344  28.70813397 28.71  Neutral   319  3344   9.539474   9.54
56 Secret Identity      Female   960  3344  28.70813397 28.71     null    94  3344   2.811005   2.81
57 Secret Identity Transgender     4  3344   0.11961722  0.12     Good  1542  3344  46.112440  46.11
58 Secret Identity Transgender     4  3344   0.11961722  0.12      Bad  1389  3344  41.537081  41.54
59 Secret Identity Transgender     4  3344   0.11961722  0.12  Neutral   319  3344   9.539474   9.54
60 Secret Identity Transgender     4  3344   0.11961722  0.12     null    94  3344   2.811005   2.81
61 Secret Identity  Non-binary     3  3344   0.08971292  0.09     Good  1542  3344  46.112440  46.11
62 Secret Identity  Non-binary     3  3344   0.08971292  0.09      Bad  1389  3344  41.537081  41.54
63 Secret Identity  Non-binary     3  3344   0.08971292  0.09  Neutral   319  3344   9.539474   9.54
64 Secret Identity  Non-binary     3  3344   0.08971292  0.09     null    94  3344   2.811005   2.81
65 Secret Identity        null    38  3344   1.13636364  1.14     Good  1542  3344  46.112440  46.11
66 Secret Identity        null    38  3344   1.13636364  1.14      Bad  1389  3344  41.537081  41.54
67 Secret Identity        null    38  3344   1.13636364  1.14  Neutral   319  3344   9.539474   9.54
68 Secret Identity        null    38  3344   1.13636364  1.14     null    94  3344   2.811005   2.81
```

```
> hc <- hclust(JoinGAnew)
> dhc <- as.dendrogram(hc)
> dhc
'dendrogram' with 2 branches and 68 members total, at height 5461.57
> specific_leaf <- dhc[[1]][[1]][[1]]
> i=0
> colLab<<-function(n){
+     if(is.leaf(n)){
+
+         a=attributes(n)
+
+
+         ligne=match(attributes(n)$label,JoinGA[,1])
+         Gender=JoinGA[ligne,3];
+         if(Gender=="Male"){col_Gender="blue"};if(Gender=="Female"){col_Gender="red"}
+
+
+         attr(n,"nodePar")<-c(a$nodePar,list(cex=1.5,lab.cex=1,pch=20,col=col_Gender,lab.font=1))
+     }
+     return(n)
+ }
```

```
> JoinGA.std=scale(JoinGA[8:10])
> d<-dist(JoinGA.std,method="euclidean")
> clust<-hclust(d,method="ward.D")
> plot(clust,cex=0.3)
> legend("topright",
+        legend = c("Male" , "Female" , "Transgender" , "Genderless" , "Non-binary","null"),
+        col = c("red", "blue" , "blue" , "red" , "Darkgreen","pink"),
+        pch = c(20,20,4,4,4,4), bty = "n",  pt.cex = 1.5, cex = 0.8 ,
+        text.col = "black", horiz = FALSE, inset = c(0, 0.1))
> rect.hclust(clust,k=3,border="pink")
>
```



7.1 (cluster dendrogram)

We use clustering analysis for determining natural groupings in multivariate data. So in this, we use agglomerative clustering which is a hierarchical clustering method. First, we standardized our selected columns in the data set. Because otherwise, we cannot prepare the model. We use hclust() with “ward.D” method, and to obtain a dissimilarity matrix we use the “Euclidean algorithm”.
There are three main clusters in this graph **(We used cutree method to obtain desired numbers of cluster which is 3)**. To clustering, we have used gender and alignment data. We

can see the alignment percentage values which are filtered with the Gender on the bottom of clusters. There is a separate percentage data point at the bottom of this dendrogram.

References

- https://www.kaggle.com/platinaz/dc-character-debut-by-year-20152020
- **Plotting Graph :-** https://www.r-graph-gallery.com/index.html
- **Hypothesis Testing**: -https://www.khanacademy.org/math/statistics-probability/significance-tests-one-sample/idea-of-significance-tests/v/simple-hypothesis-testing
- **Normal Distribution: -** https://stat.ethz.ch/R-manual/R-devel/library/stats/html/Normal.html

Individual Contributions

Name	Index Number	Contribution
W.P Pallewatta	18001149	• Observation about data set, Clustering Analysis • Plotting graphs with grouped data • Hypothesis Overview analyzation
M.H.D.S Jayalath	18000703	• Clustering Analysis • Residual Plot Creating
K.K Samaraweera	18001459	• Observation about Data set, • Relationship between variables (correlation, regression line, residual plot)
T.T Wattuhewa	18001858	• Introduction • Plot the multivariate data • Hypothesis Testing
D.J.Y.W Gamage	18000568	• Introduction • Plot the multivariate data

		• [Hypothesis Testing]

Thank You!