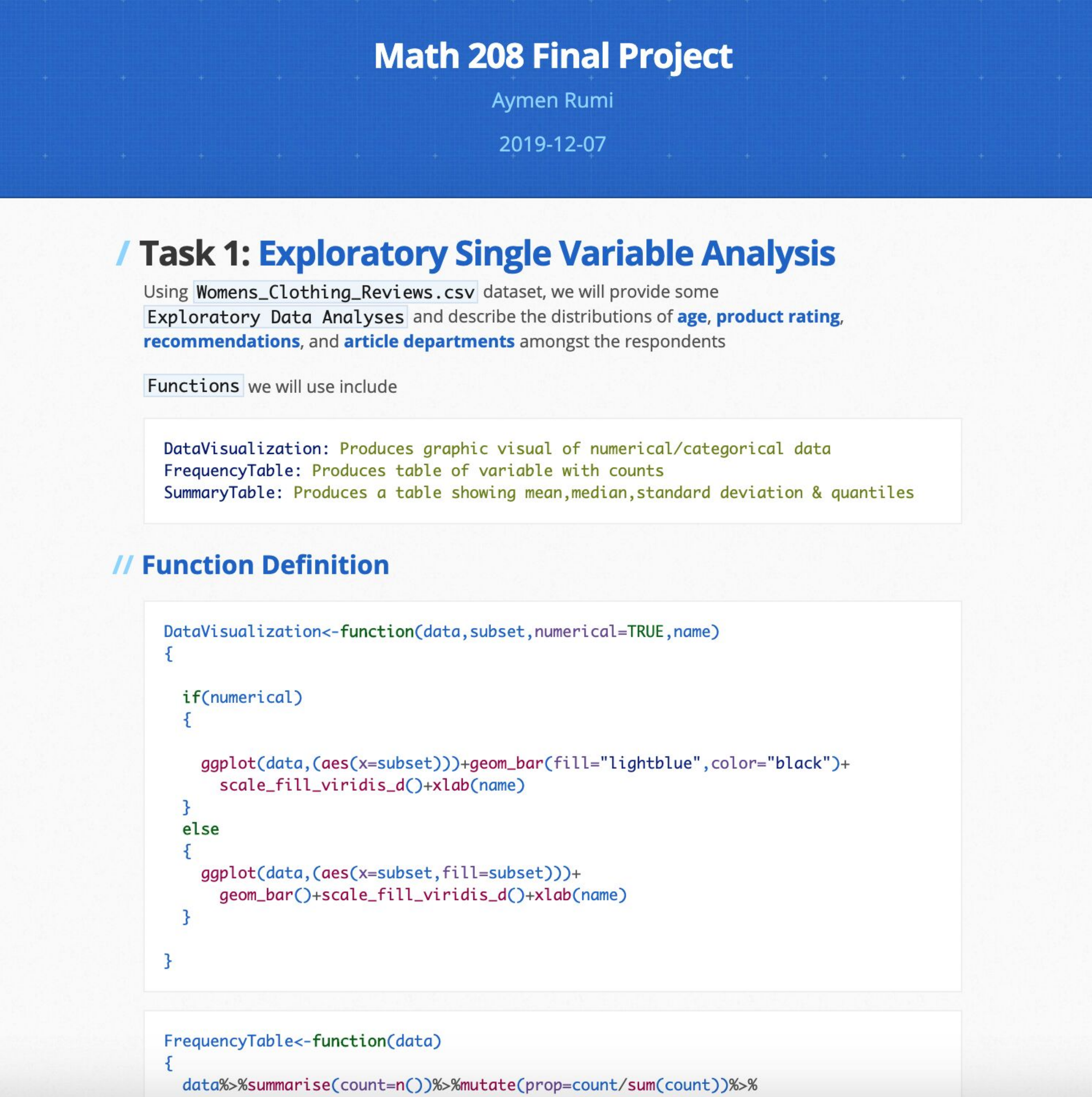


Math 208 Final Project

Aymen Rumi

2019-12-07

/ Task 1: Exploratory Single Variable Analysis

Using `Womens_Clothing_Reviews.csv` dataset, we will provide some `Exploratory Data Analyses` and describe the distributions of **age**, **product rating**, **recommendations**, and **article departments** amongst the respondents

`Functions` we will use include

```
DataVisualization: Produces graphic visual of numerical/categorical data
FrequencyTable: Produces table of variable with counts
SummaryTable: Produces a table showing mean,median,standard deviation & quantiles
```

// Function Definition

```
DataVisualization<-function(data,subset,numerical=TRUE,name)
{

  if(numerical)
  {

    ggplot(data,(aes(x=subset)))+geom_bar(fill="lightblue",color="black")+
      scale_fill_viridis_d()+xlab(name)
  }
  else
  {
    ggplot(data,(aes(x=subset,fill=subset)))+
      geom_bar()+scale_fill_viridis_d()+xlab(name)
  }

}
```

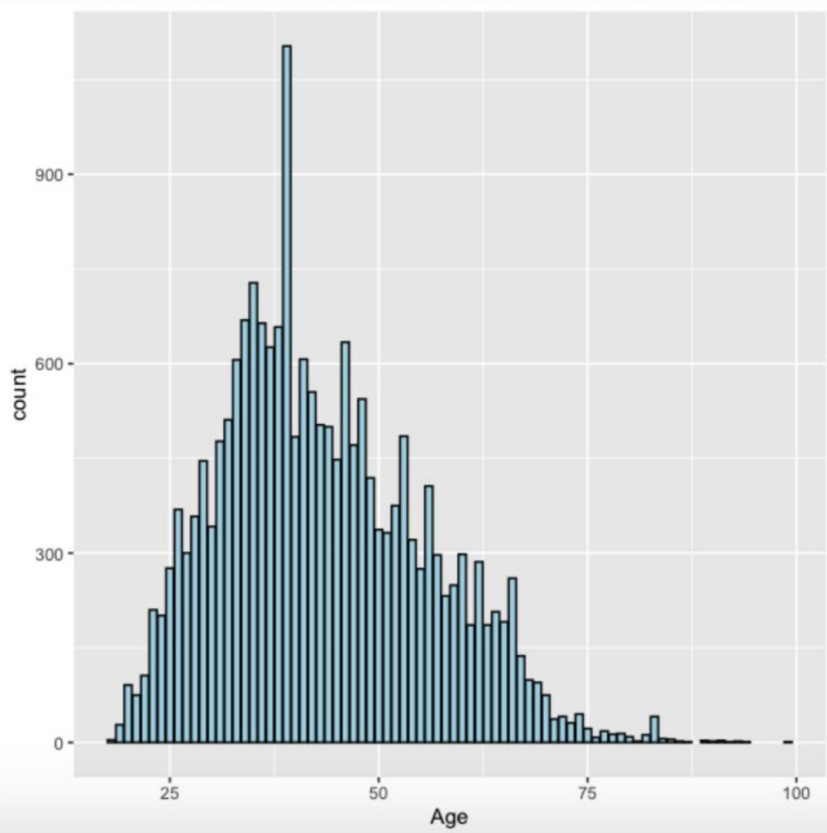
```
FrequencyTable<-function(data)
{
  data%>%summarise(count=n())%>%mutate(prop=count/sum(count))%>%
    arrange(desc(count))%>%kable()
}
```

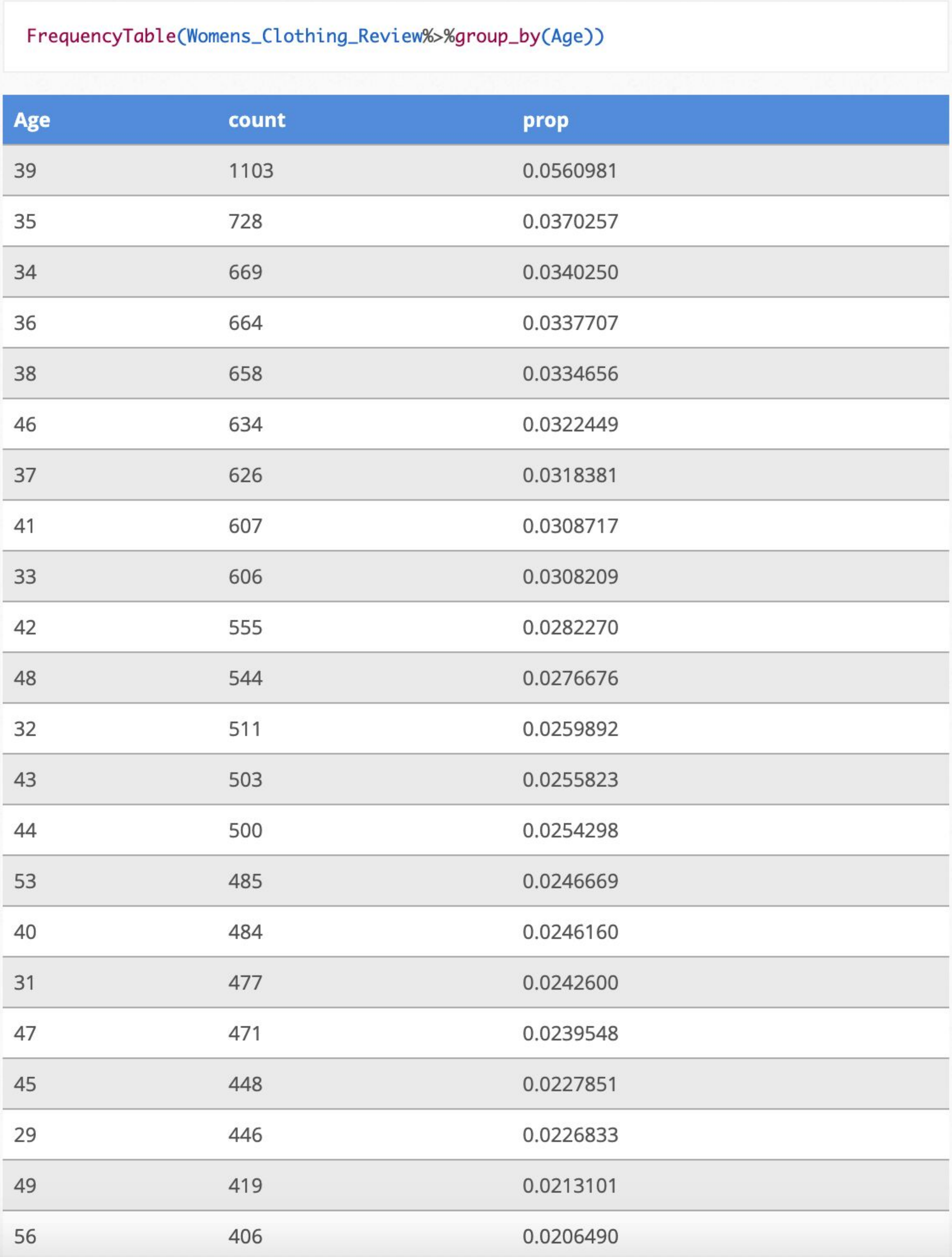
```
SummaryTable<-function(data,subset)
{
  data%>%summarise(Ave=mean(subset),
                   Med=median(subset),
                   '25%ile'=quantile(subset,0.25),
                   '75%ile'=quantile(subset,0.75),
                   Std=sd(subset)
                   )%>%kable()
}
```

// Variable Analysis: Age

The figures below provide the `Density Plot`, `Ordered Count/Proportion Table`, and `Summary Table` for **age**

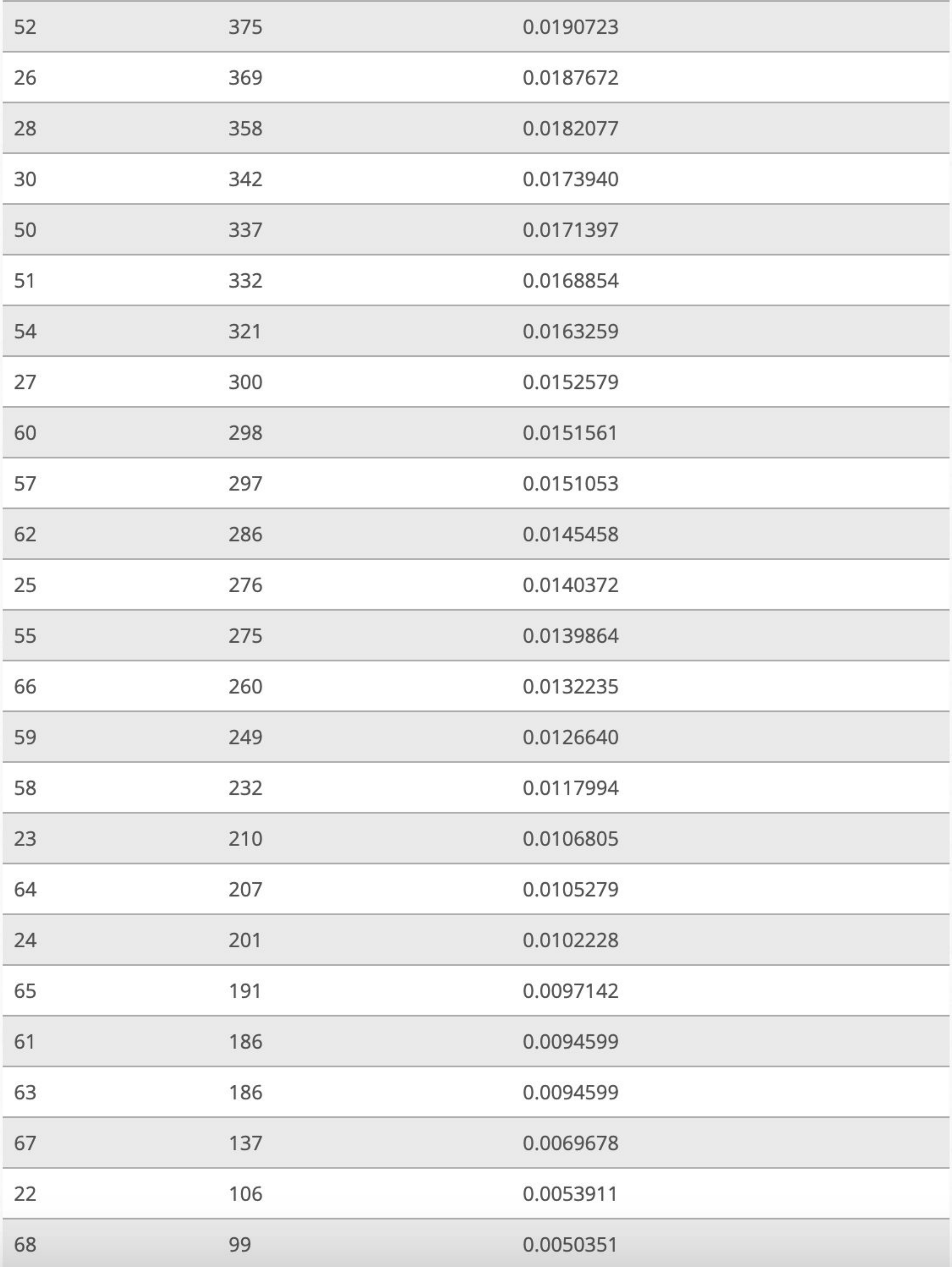
```
DataVisualization(Womens_Clothing_Review,with(Womens_Clothing_Review,Age),name="Age")
```



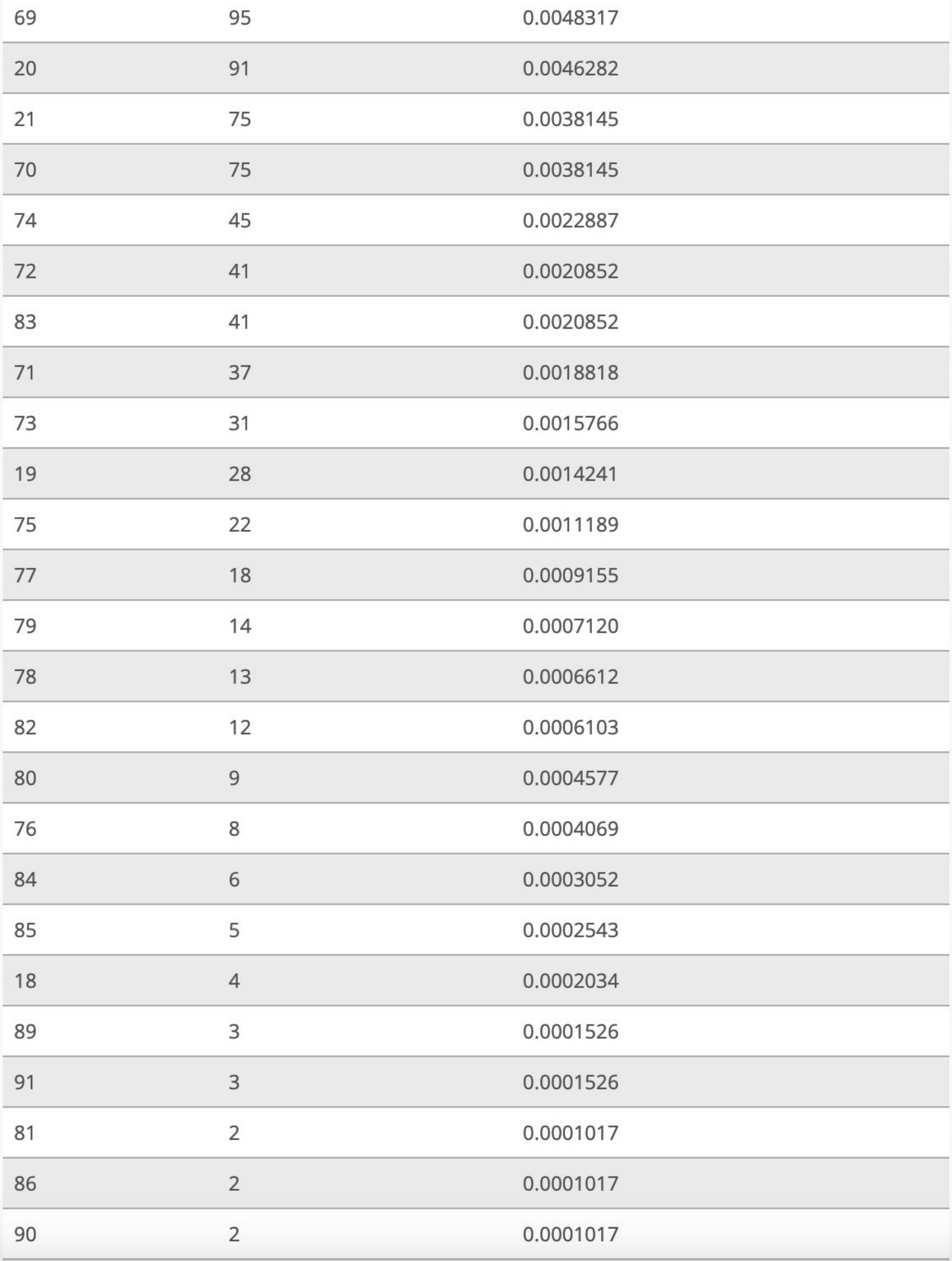


```
FrequencyTable(Womens_Clothing_Review%>%group_by(Age))
```

Age	count	prop
39	1103	0.0560981
35	728	0.0370257
34	669	0.0340250
36	664	0.0337707
38	658	0.0334656
46	634	0.0322449
37	626	0.0318381
41	607	0.0308717
33	606	0.0308209
42	555	0.0282270
48	544	0.0276676
32	511	0.0259892
43	503	0.0255823
44	500	0.0254298
53	485	0.0246669
40	484	0.0246160
31	477	0.0242600
47	471	0.0239548
45	448	0.0227851
29	446	0.0226833
49	419	0.0213101
56	406	0.0206490



52	375	0.0190723
26	369	0.0187672
28	358	0.0182077
30	342	0.0173940
50	337	0.0171397
51	332	0.0168854
54	321	0.0163259
27	300	0.0152579
60	298	0.0151561
57	297	0.0151053
62	286	0.0145458
25	276	0.0140372
55	275	0.0139864
66	260	0.0132235
59	249	0.0126640
58	232	0.0117994
23	210	0.0106805
64	207	0.0105279
24	201	0.0102228
65	191	0.0097142
61	186	0.0094599
63	186	0.0094599
67	137	0.0069678
22	106	0.0053911
68	99	0.0050351



69	95	0.0048317
20	91	0.0046282
21	75	0.0038145
70	75	0.0038145
74	45	0.0022887
72	41	0.0020852
83	41	0.0020852
71	37	0.0018818
73	31	0.0015766
19	28	0.0014241
75	22	0.0011189
77	18	0.0009155
79	14	0.0007120
78	13	0.0006612
82	12	0.0006103
80	9	0.0004577
76	8	0.0004069
84	6	0.0003052
85	5	0.0002543
18	4	0.0002034
89	3	0.0001526
91	3	0.0001526
81	2	0.0001017
86	2	0.0001017
90	2	0.0001017

93	2	0.0001017
87	1	0.0000509
92	1	0.0000509
94	1	0.0000509
99	1	0.0000509

```
SummaryTable(Womens_Clothing_Review,with(Womens_Clothing_Review,Age))
```

Ave	Med	25%ile	75%ile	Std
43.26081	41	34	52	12.25812

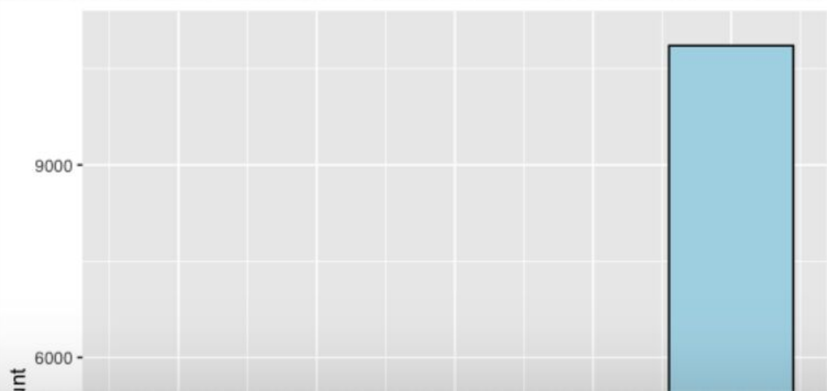
`Analysis`

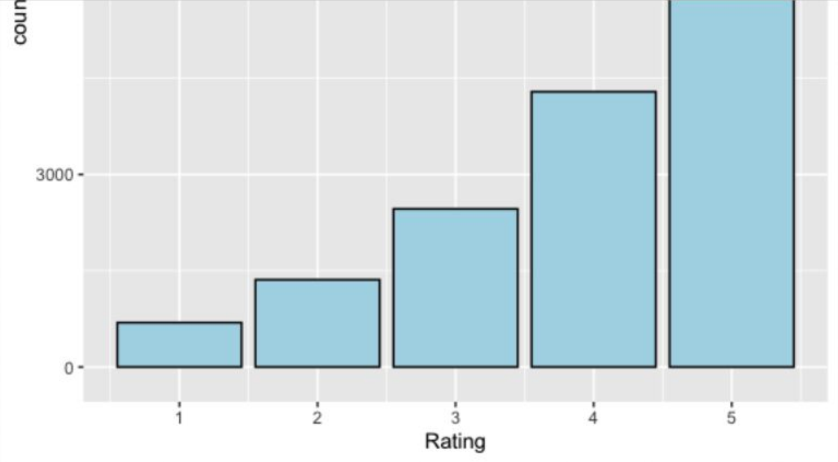
```
We can see that the average age is approximately 43 with
a spread of 12, the density has a high peak for age
39(1103 people) which is 375 more than the next highest
age group of 35, the 25th and 75th quartile are between
34 and 53
```

// Variable Analysis: Rating

The figures below provide the `Density Plot` and `Ordered Count/Proportion Table` for **rating**

```
DataVisualization(Womens_Clothing_Review,
                  with(Womens_Clothing_Review,Rating),name="Rating")
```





```
FrequencyTable(Womens_Clothing_Review%>%group_by(Rating))
```

Rating	count	prop
5	10858	0.5522327
4	4289	0.2181365
3	2464	0.1253179
2	1360	0.0691690
1	691	0.0351439

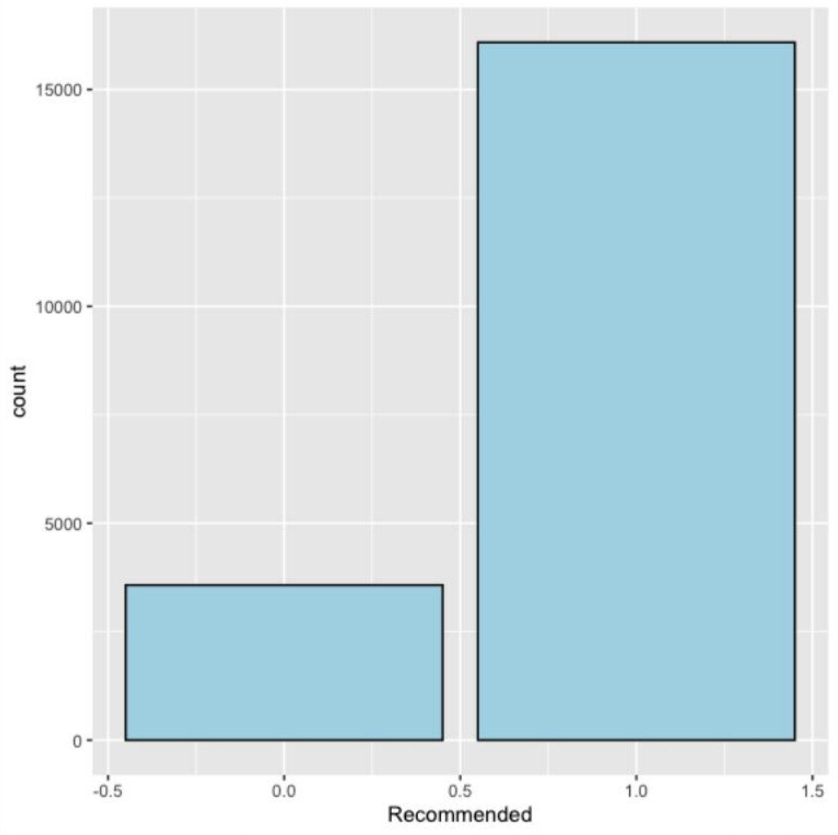
`Analysis`

```
Looks like more a little more than half of reviews are positive with 5 stars,
and the rest combine for the other half with 4 taking approximately 20%
3 taking 12% and the rest for 1&2, overall very good ratings
```

// Variable Analysis: Recommendations

The figures below provide the `Density Plot` and `Ordered Count/Proportion Table` for **recommendations**

```
DataVisualization(Womens_Clothing_Review,
                  with(Womens_Clothing_Review,Recommended),name="Recommended")
```



```
FrequencyTable(Womens_Clothing_Review%>%group_by(Recommended))
```

Recommended	count	prop
1	16087	0.8181772
0	3575	0.1818228

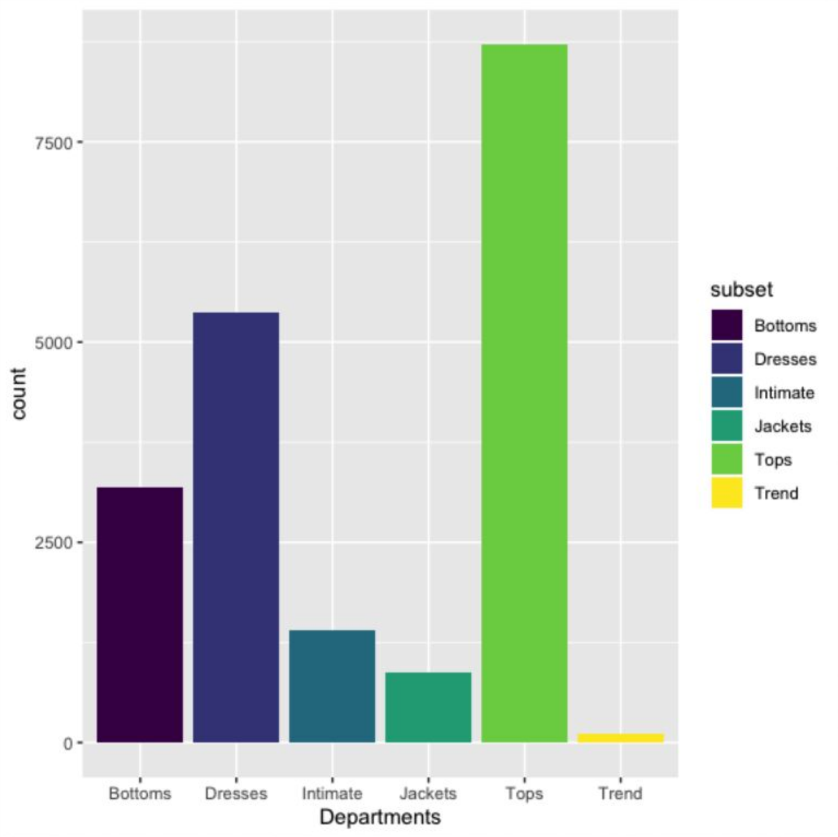
`Analysis`

```
There is a 80% positive recommendation and 20% negative,
overall high recommendation rating
```

// Variable Analysis: Departments

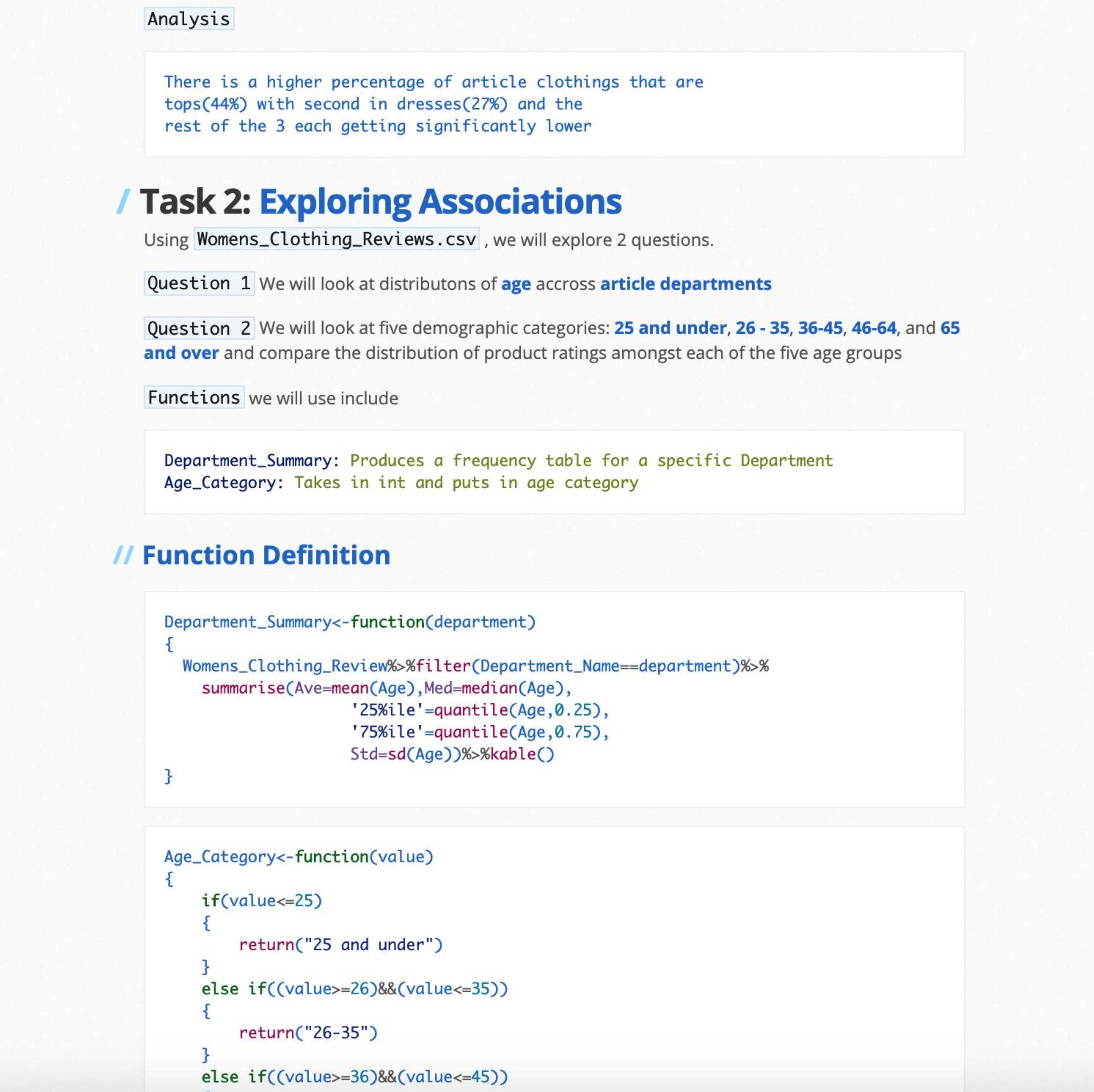
The figures below provide the `Density Plot` and `Ordered Count/Proportion Table` for **recommendations**

```
DataVisualization(Womens_Clothing_Review,with(Womens_Clothing_Review,
                                              Department_Name),FALSE,"Departments")
```



```
FrequencyTable(Womens_Clothing_Review%>%
                 group_by(Department_Name))
```

Department_Name	count	prop
Tops	8713	0.4431390
Dresses	5371	0.2731665
Bottoms	3184	0.1619367
Intimate	1408	0.0716102
Jackets	879	0.0447055
Trend	107	0.0054420



`Analysis`

```
There is a higher percentage of article clothings that are
tops(44%) with second in dresses(27%) and the
rest of the 3 each getting significantly lower
```

Task 2: Exploring Associations

Using `Womens_Clothing_Reviews.csv` , we will explore 2 questions.

`Question 1` We will look at distributons of **age** accross **article departments**

`Question 2` We will look at five demographic categories: **25 and under**, **26 - 35**, **36-45**, **46-64**, and **65 and over** and compare the distribution of product ratings amongst each of the five age groups

`Functions` we will use include

```
Department_Summary: Produces a frequency table for a specific Department
Age_Category: Takes in int and puts in age category
```

Function Definition

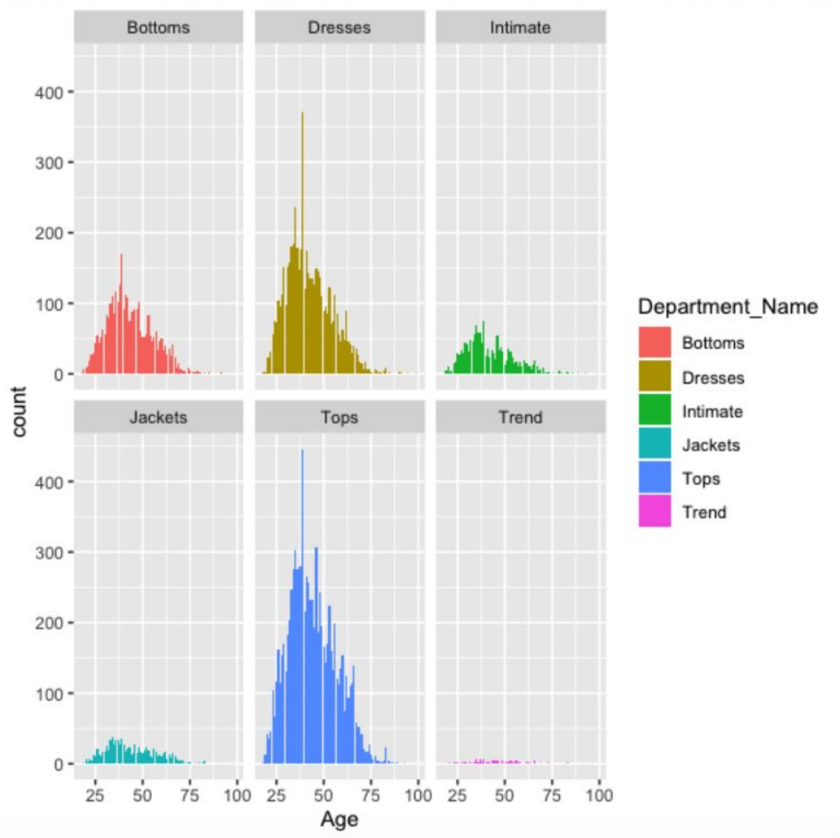
```
Department_Summary<-function(department)
{
  Womens_Clothing_Review%>%filter(Department_Name==department)%>%
    summarise(Ave=mean(Age),Med=median(Age),
                      '25%ile'=quantile(Age,0.25),
                      '75%ile'=quantile(Age,0.75),
                      Std=sd(Age))%>%kable()
}
```

```
Age_Category<-function(value)
{
    if(value<=25)
    {
        return("25 and under")
    }
    else if((value>=26)&&(value<=35))
    {
        return("26-35")
    }
    else if((value>=36)&&(value<=45))
    {
        return ("36-45")
    }
    else if((value>=36)&&(value<=64))
    {
        return("46-64")
    }
    else
    {
        return("65 and over")
    }
}
```

// Question 1:

`Data Visualization`

```
ggplot(Womens_Clothing_Review,aes(Age,fill=Department_Name))+
  geom_bar()+facet_wrap(~Department_Name)
```



`Bottoms: Age Distribution Table`

```
Department_Summary("Bottoms")
```

Ave	Med	25%ile	75%ile	Std
43.18467	41	35	51	11.76209

`Intimate: Age Distribution Table`

```
Department_Summary("Intimate")
```

Ave	Med	25%ile	75%ile	Std
41.63352	39	33	49	12.28109

`Jackets: Age Distribution Table`

```
Department_Summary("Jackets")
```

Ave	Med	25%ile	75%ile	Std
43.96132	42	34	53	12.95756

`Tops: Age Distribution Table`

```
Department_Summary("Tops")
```

Ave	Med	25%ile	75%ile	Std
44.12579	42	35	53	12.50254

`Trends: Age Distribution Table`

```
Department_Summary("Trend")
```

Ave	Med	25%ile	75%ile	Std
44.34579	43	36.5	53	12.21899

`Analysis`

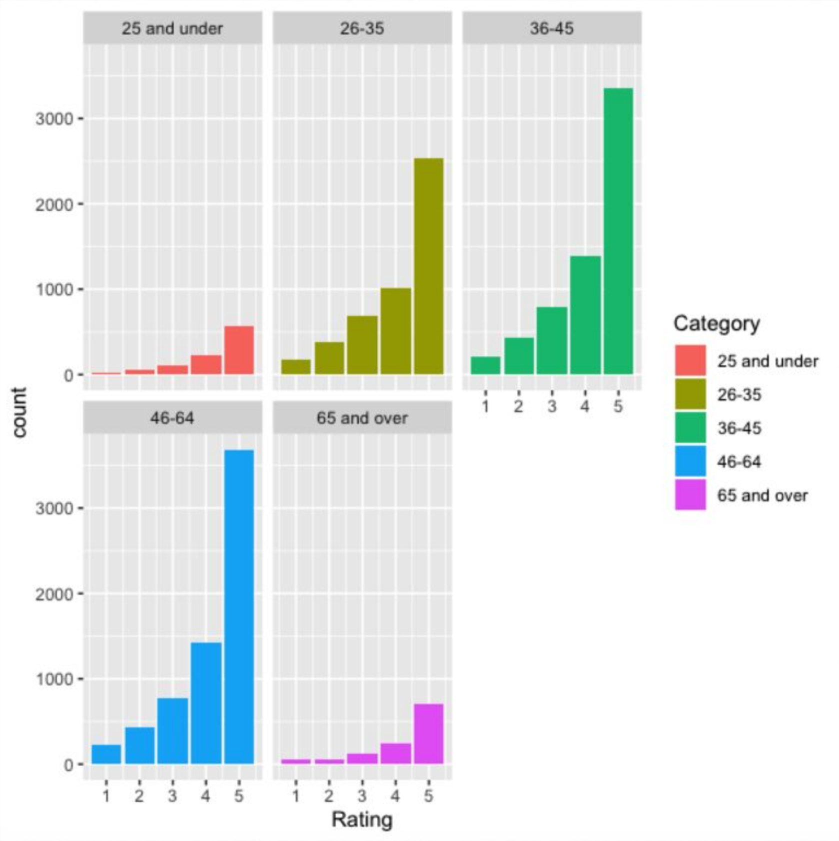
```
The distribution as in the spread of age seems to look relatively the same across
all departments but the densities differ greatly,the density of
people rating products are in order of Tops, Dresses, Jackets,
Intimate,Jackets, Trends
```

// Question 2:

`Data Visualization`

```
Womens_Clothing_Review<-Womens_Clothing_Review%>%
  mutate(Category=map_chr(Age,Age_Category))


ggplot(Womens_Clothing_Review,aes(x=Rating,fill=Category))+
  geom_bar()+facet_wrap(~Category)
```



`Analysis`

```
Again the distribution as in the spread of rating seems to look
relatively the same across all age groups but the densities
differ greatly due to the fact that there is more counts of
ratings from certain groups, groups such as 46-64 and 36-45
age group have a vast amount of ratings and seem to be most
enthusiastic about their company’s products.
```

// Task 3: Ten Most Popular Products

Using `Womens_Clothing_Reviews.csv` dataset, we will compile a list of their **ten most popular products** based on **Recommendations** (with each product indicated by **ID number**)

We will use `Wilson’s Lower Confidence Limit` computed via:

n = the `number` of respondents who rated that product (positively or negatively)

p=the `proportion` of respondents who positively recommended a certain product

$$a = \frac{1.96^2}{n}$$

$$b = \frac{p(1-p)}{n}$$

$$c = \frac{a}{2n}$$

$$WLCL = \frac{p+a-1.96\sqrt{b+c}}{1+2a}$$

`Functions` we will use include

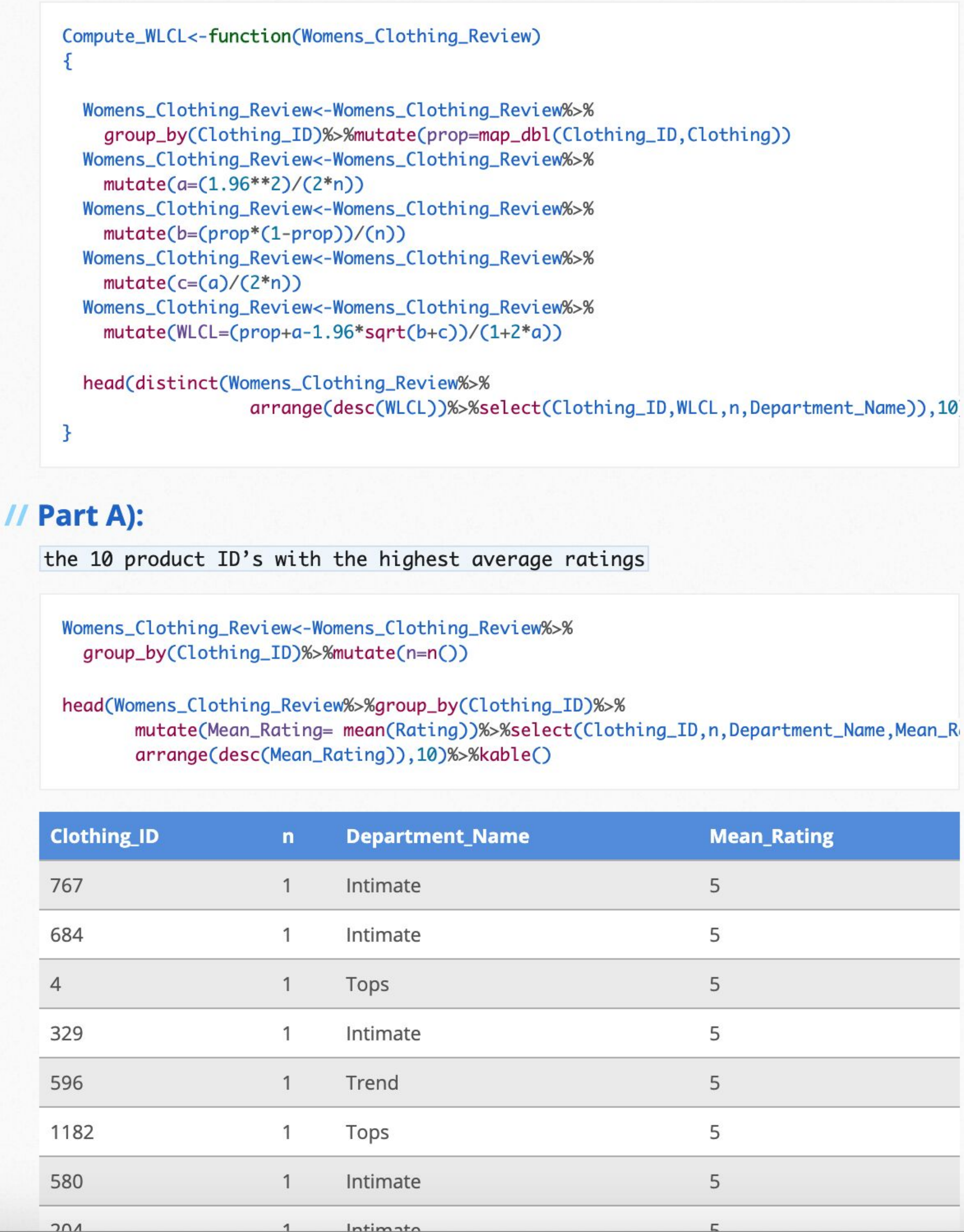
```
Clothing: returns the proportion of positively recommended ratings for given ID
Compute_WLCL: Given a dataset will compute its WLCL and output it
```

// Function Definition

```
Clothing<-function(ID)
{

    max((Womens_Clothing_Review%>%filter(Clothing_ID%in%ID)%>%
           group_by(Recommended)%>%summarise(count=n()))%>%
          mutate(prop=count/sum(count))%>%select(prop))

}
```



```
Compute_WLCL<-function(Womens_Clothing_Review)
{

  Womens_Clothing_Review<-Womens_Clothing_Review%>%
    group_by(Clothing_ID)%>%mutate(prop=map_dbl(Clothing_ID,Clothing))
  Womens_Clothing_Review<-Womens_Clothing_Review%>%
    mutate(a=(1.96**2)/(2*n))
  Womens_Clothing_Review<-Womens_Clothing_Review%>%
    mutate(b=(prop*(1-prop))/(n))
  Womens_Clothing_Review<-Womens_Clothing_Review%>%
    mutate(c=(a)/(2*n))
  Womens_Clothing_Review<-Womens_Clothing_Review%>%
    mutate(WLCL=(prop+a-1.96*sqrt(b+c))/(1+2*a))

  head(distinct(Womens_Clothing_Review%>%
                  arrange(desc(WLCL))%>%select(Clothing_ID,WLCL,n,Department_Name)),10
}
```

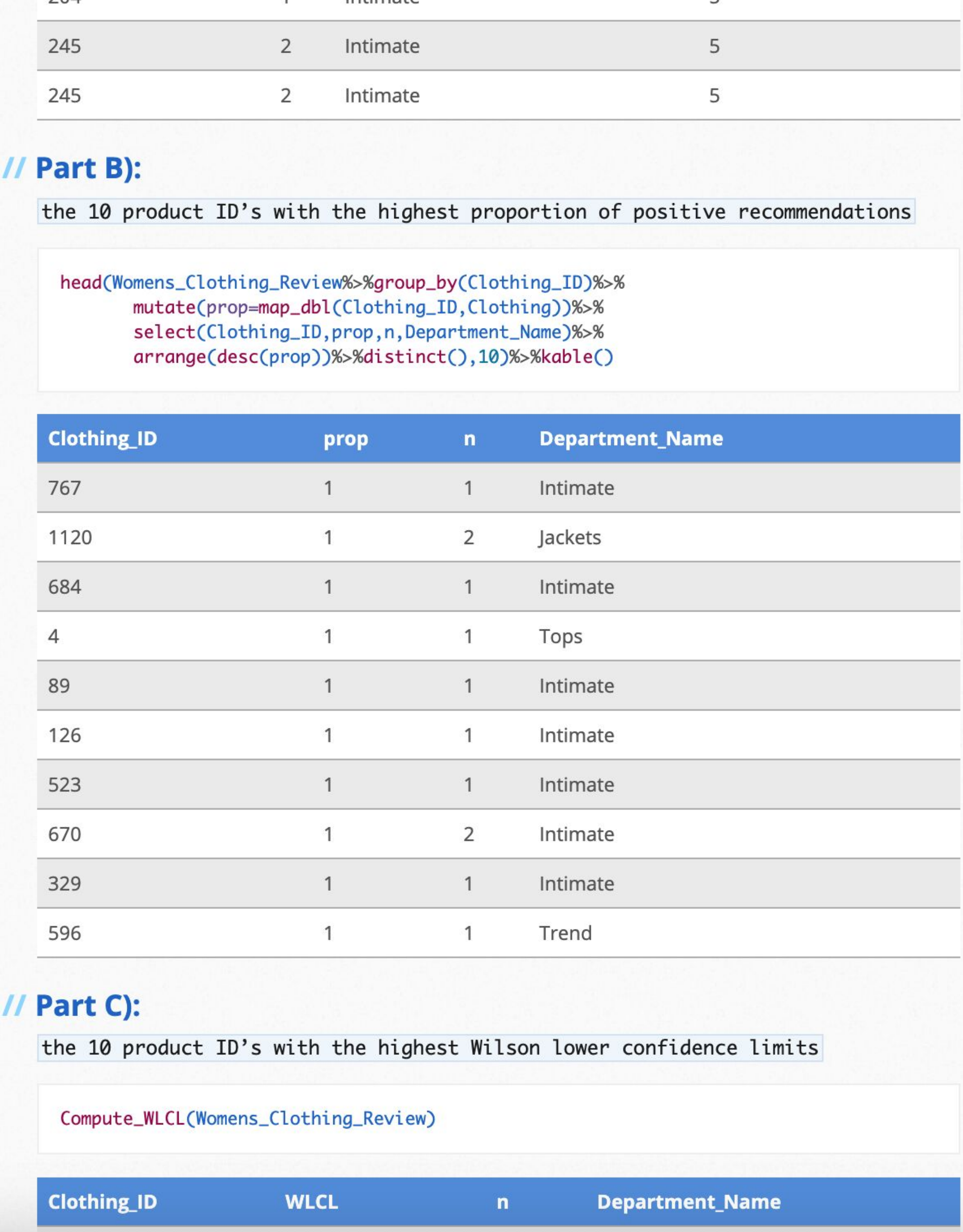
// Part A):

`the 10 product ID’s with the highest average ratings`

```
Womens_Clothing_Review<-Womens_Clothing_Review%>%
  group_by(Clothing_ID)%>%mutate(n=n())

head(Womens_Clothing_Review%>%group_by(Clothing_ID)%>%
       mutate(Mean_Rating= mean(Rating))%>%select(Clothing_ID,n,Department_Name,Mean_R
       arrange(desc(Mean_Rating)),10)%>%kable()
```

Clothing_ID	n	Department_Name	Mean_Rating
767	1	Intimate	5
684	1	Intimate	5
4	1	Tops	5
329	1	Intimate	5
596	1	Trend	5
1182	1	Tops	5
580	1	Intimate	5
204	1	Intimate	5



245	2	Intimate	5
245	2	Intimate	5

// Part B):

`the 10 product ID’s with the highest proportion of positive recommendations`

```
head(Womens_Clothing_Review%>%group_by(Clothing_ID)%>%
       mutate(prop=map_dbl(Clothing_ID,Clothing))%>%
       select(Clothing_ID,prop,n,Department_Name)%>%
       arrange(desc(prop))%>%distinct(),10)%>%kable()
```

Clothing_ID	prop	n	Department_Name
767	1	1	Intimate
1120	1	2	Jackets
684	1	1	Intimate
4	1	1	Tops
89	1	1	Intimate
126	1	1	Intimate
523	1	1	Intimate
670	1	2	Intimate
329	1	1	Intimate
596	1	1	Trend

// Part C):

`the 10 product ID’s with the highest Wilson lower confidence limits`

```
Compute_WLCL(Womens_Clothing_Review)
```

Clothing_ID	WLCL	n	Department_Name

523	1	1	Intimate
670	1	2	Intimate
329	1	1	Intimate
596	1	1	Trend

// Part C):

`the 10 product ID’s with the highest Wilson lower confidence limits`

```
Compute_WLCL(Womens_Clothing_Review)
```

Clothing_ID	WLCL	n	Department_Name
964	0.8728595	65	Jackets
834	0.8688158	125	Tops
1123	0.8668035	25	Jackets
520	0.8620194	24	Intimate
1008	0.8610012	163	Bottoms
1025	0.8500173	100	Bottoms
839	0.8454422	43	Tops
984	0.8434221	144	Jackets
1022	0.8339168	172	Bottoms
1033	0.8309123	190	Bottoms

`Analysis`

```
I think the list that best represents the products which are most popular are,
List C for WLCL because the lists shown in A and B can be deceiving as the
number of counts can be low and this matters for showing an indication of popularity a
i think that WLCL captures that well, as opposed to showing just rating and positive
recommendation proportions it measures the popularity
```