In [2]:
```
#import packages
library(tidyverse)
library(ggplot2)
library(grid)
library(GGally)
library(gridExtra)
```

In [3]:
```
#import data
df <- read.csv("train.csv")
```

In [4]:
```
#examine data
head(df)
```

	Patient_Id	Patient_Age	Genes_in_mother_s_side	Inherited_from_father	Maternal_gene	Paternal
	<chr>	<int>	<chr>	<chr>	<chr>	
1	PID0x8ce3	11	No	No	Yes	
2	PID0x8660	4	No	Yes	Yes	
3	PID0x74ab	1	Yes	Yes	No	
4	PID0x7678	6	Yes	No	Yes	
5	PID0x952d	10	Yes	Yes	Yes	
6	PID0x6d89	6	No	Yes	Yes	

In [5]:
```
#dataframe structure
str(df)
```

```
'data.frame':   6706 obs. of  45 variables:
 $ Patient_Id                                     : chr  "PID0x8ce3" "PID0x8660"
"PID0x74ab" "PID0x7678" ...
 $ Patient_Age                                    : int  11 4 1 6 10 6 10 4 8 1
...
 $ Genes_in_mother_s_side                         : chr  "No" "No" "Yes" "Yes"
...
 $ Inherited_from_father                          : chr  "No" "Yes" "Yes" "No"
...
 $ Maternal_gene                                  : chr  "Yes" "Yes" "No" "Yes"
...
 $ Paternal_gene                                  : chr  "No" "Yes" "No" "No" ...
 $ Blood_cell_count__mcL_                         : num  5.21 4.75 4.61 4.62 4.75
...
 $ Patient_First_Name                             : chr  "Willie" "John" "Eric"
"Ruth" ...
 $ Family_Name                                    : chr  "Camacho" "Sandoval" "Ha
rness" "Homza" ...
 $ Father_s_name                                  : chr  "Tr" "Gregori" "Mano" "C
esareo" ...
 $ Mother_s_age                                   : int  45 44 50 41 40 36 30 49
18 38 ...
 $ Father_s_age                                   : int  44 42 56 20 57 48 42 28
31 61 ...
 $ Institute_Name                                 : chr  "Lemuel Shattuck Hospita
l" "Shriners Burns Institute" "Not applicable" "Not applicable" ...
 $ Location_of_Institute                          : chr  "125 NASHUA ST\nCENTRAL,
MA 02114\n(42.36764789068138, -71.06564730220646)" "1200 Centre St\nRoslindale, MA
02131\n(42.29738386053219, -71.13150465441208)" "-" "-" ...
 $ Status                                         : chr  "Alive" "Alive" "Decease
d" "Alive" ...
 $ Respiratory_Rate__breaths_min_                 : chr  "Tachypnea" "Tachypnea"
"Normal (30-60)" "Tachypnea" ...
 $ Heart_Rate__rates_min                          : chr  "Tachycardia" "Tachycard
ia" "Tachycardia" "Tachycardia" ...
 $ Test_1                                         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Test_2                                         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Test_3                                         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Test_4                                         : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Test_5                                         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Parental_consent                               : chr  "Yes" "Yes" "Yes" "Yes"
...
 $ Follow.up                                      : chr  "Low" "Low" "High" "Hig
h" ...
 $ Gender                                         : chr  "Male" "Male" "Male" "Ma
le" ...
 $ Birth_asphyxia                                 : chr  "Yes" "No" "Not availabl
e" "No" ...
 $ Autopsy_shows_birth_defect__if_applicable_     : chr  "Not applicable" "Not ap
plicable" "None" "Not applicable" ...
 $ Place_of_birth                                 : chr  "Institute" "Institute"
"Home" "Home" ...
 $ Folic_acid_details__peri.conceptional_         : chr  "Yes" "Yes" "Yes" "Yes"
...
 $ H_O_serious_maternal_illness                   : chr  "Yes" "No" "No" "Yes"
...
```

```
 $ H_O_radiation_exposure__x.ray_                  : chr  "No" "No" "Yes" "Yes"
...
 $ H_O_substance_abuse                             : chr  "No" "No" "Not applicabl
e" "-" ...
 $ Assisted_conception_IVF_ART                     : chr  "No" "Yes" "Yes" "No"
...
 $ History_of_anomalies_in_previous_pregnancies    : chr  "Yes" "Yes" "Yes" "No"
...
 $ No._of_previous_abortion                        : int  0 1 0 3 3 1 0 4 0 4 ...
 $ Birth_defects                                   : chr  "Multiple" "Multiple" "S
ingular" "Multiple" ...
 $ White_Blood_cell_count__thousand_per_microliter_: num  6.67 6.4 8 3 9.38 ...
 $ Blood_test_result                               : chr  "slightly abnormal" "abn
ormal" "slightly abnormal" "slightly abnormal" ...
 $ Symptom_1                                       : int  1 0 1 1 1 1 1 1 1 0 ...
 $ Symptom_2                                       : int  1 0 1 0 1 0 1 0 0 1 ...
 $ Symptom_3                                       : int  1 1 0 1 0 0 1 0 1 0 ...
 $ Symptom_4                                       : int  0 1 1 0 0 0 1 1 0 1 ...
 $ Symptom_5                                       : int  1 1 0 1 0 0 1 0 0 0 ...
 $ Genetic_Disorder                                : chr  "Mitochondrial genetic i
nheritance disorders" "Multifactorial genetic inheritance disorders" "Mitochondria
l genetic inheritance disorders" "Mitochondrial genetic inheritance disorders" ...
 $ Disorder_Subclass                               : chr  "Leigh syndrome" "Diabet
es" "Leigh syndrome" "Leigh syndrome" ...
```

In [6]:
```r
#columns not necessary for further calculations
drop <- c("Patient_Id","Patient_First_Name","Family_Name"
          ,"Father_s_name","Institute_Name","Location_of_Institute","Parental_conse
df = df[,!(names(df) %in% drop)]
```

In [7]:
```r
#summary statistics
summary(df)
```

```
 Patient_Age     Genes_in_mother_s_side Inherited_from_father
Min.   : 0.000   Length:6706            Length:6706
1st Qu.: 3.000   Class :character       Class :character
Median : 7.000   Mode  :character       Mode  :character
Mean   : 6.916
3rd Qu.:11.000
Max.   :14.000
Maternal_gene      Paternal_gene      Blood_cell_count__mcL_  Mother_s_age
Length:6706        Length:6706        Min.   :4.146          Min.   :18.00
Class :character   Class :character   1st Qu.:4.767          1st Qu.:26.00
Mode  :character   Mode  :character   Median :4.900          Median :35.00
                                      Mean   :4.901          Mean   :34.64
                                      3rd Qu.:5.036          3rd Qu.:43.00
                                      Max.   :5.610          Max.   :51.00
 Father_s_age      Status          Respiratory_Rate__breaths_min_
Min.   :20.00   Length:6706        Length:6706
1st Qu.:31.00   Class :character   Class :character
Median :42.00   Mode  :character   Mode  :character
Mean   :41.99
3rd Qu.:53.00
Max.   :64.00
Heart_Rate__rates_min     Test_1      Test_2      Test_3      Test_4
Length:6706           Min.   :0   Min.   :0   Min.   :0   Min.   :1
Class :character      1st Qu.:0   1st Qu.:0   1st Qu.:0   1st Qu.:1
Mode  :character      Median :0   Median :0   Median :0   Median :1
                      Mean   :0   Mean   :0   Mean   :0   Mean   :1
                      3rd Qu.:0   3rd Qu.:0   3rd Qu.:0   3rd Qu.:1
                      Max.   :0   Max.   :0   Max.   :0   Max.   :1
    Test_5   Follow.up            Gender          Birth_asphyxia
Min.   :0   Length:6706        Length:6706        Length:6706
1st Qu.:0   Class :character   Class :character   Class :character
Median :0   Mode  :character   Mode  :character   Mode  :character
Mean   :0
3rd Qu.:0
Max.   :0
Autopsy_shows_birth_defect__if_applicable_
Length:6706
Class :character
Mode  :character



Folic_acid_details__peri.conceptional_ H_O_serious_maternal_illness
Length:6706                            Length:6706
Class :character                       Class :character
Mode  :character                       Mode  :character



H_O_radiation_exposure__x.ray_ H_O_substance_abuse Assisted_conception_IVF_ART
Length:6706                    Length:6706         Length:6706
Class :character               Class :character    Class :character
Mode  :character               Mode  :character    Mode  :character
```

```
 History_of_anomalies_in_previous_pregnancies No._of_previous_abortion
 Length:6706                                  Min.   :0.000
 Class :character                             1st Qu.:1.000
 Mode  :character                             Median :2.000
                                              Mean   :1.999
                                              3rd Qu.:3.000
                                              Max.   :4.000
 Birth_defects      White_Blood_cell_count__thousand_per_microliter_
 Length:6706        Min.   : 3.000
 Class :character   1st Qu.: 5.355
 Mode  :character   Median : 7.367
                    Mean   : 7.419
                    3rd Qu.: 9.439
                    Max.   :12.000
 Blood_test_result    Symptom_1       Symptom_2       Symptom_3
 Length:6706        Min.   :0.000   Min.   :0.0000   Min.   :0.0000
 Class :character   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000
 Mode  :character   Median :1.000   Median :1.0000   Median :1.0000
                    Mean   :0.589   Mean   :0.5495   Mean   :0.5391
                    3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000
                    Max.   :1.000   Max.   :1.0000   Max.   :1.0000
   Symptom_4        Symptom_5      Genetic_Disorder   Disorder_Subclass
 Min.   :0.0000   Min.   :0.0000   Length:6706        Length:6706
 1st Qu.:0.0000   1st Qu.:0.0000   Class :character   Class :character
 Median :1.0000   Median :0.0000   Mode  :character   Mode  :character
 Mean   :0.5013   Mean   :0.4705
 3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.0000
```

In [8]:
```
#better summary statistics for numeric columns
numeric_columns <- sapply(df, is.numeric)
sapply(df[, numeric_columns], function(x) summary(x))
```

	Patient_Age	Blood_cell_count__mcL_	Mother_s_age	Father_s_age	Test_1	Test_2	Test_3
Min.	0.000000	4.146230	18.00000	20.00000	0	0	0
1st Qu.	3.000000	4.766541	26.00000	31.00000	0	0	0
Median	7.000000	4.899961	35.00000	42.00000	0	0	0
Mean	6.915896	4.900562	34.64002	41.98539	0	0	0
3rd Qu.	11.000000	5.036320	43.00000	53.00000	0	0	0
Max.	14.000000	5.609829	51.00000	64.00000	0	0	0

In [9]:
```
#columns with no variance
drop <- c("Test_1","Test_2","Test_3","Test_4","Test_5")
df = df[,!(names(df) %in% drop)]
```

In [10]:
```
head(df)
tail(df)
```

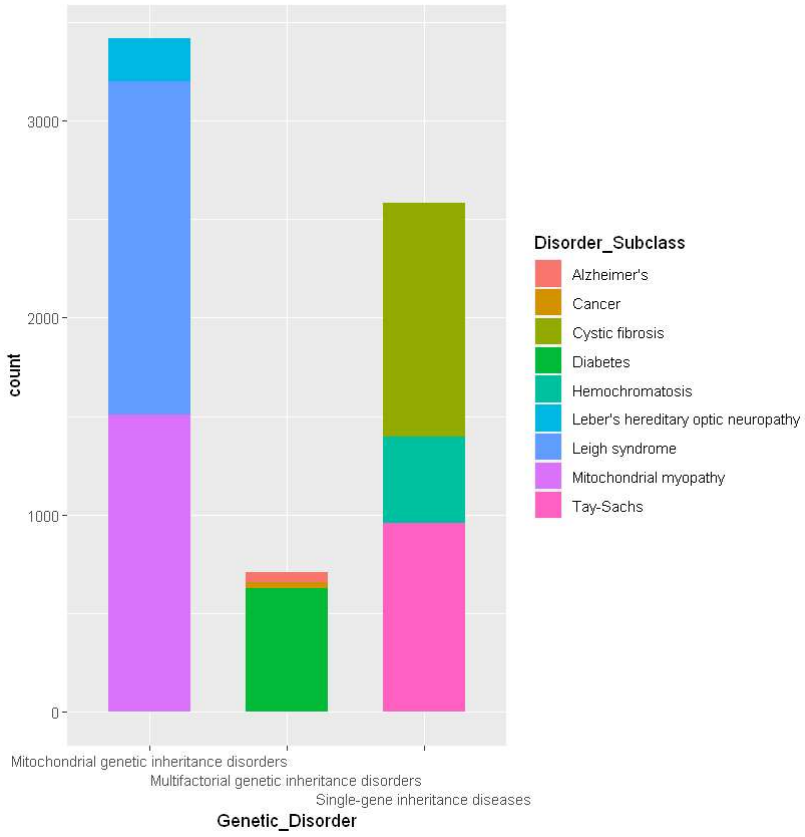
	Patient_Age <int>	Genes_in_mother_s_side <chr>	Inherited_from_father <chr>	Maternal_gene <chr>	Paternal_gene <chr>	Bloo
1	11	No	No	Yes	No	
2	4	No	Yes	Yes	Yes	
3	1	Yes	Yes	No	No	
4	6	Yes	No	Yes	No	
5	10	Yes	Yes	Yes	No	
6	6	No	Yes	Yes	Yes	

	Patient_Age	Genes_in_mother_s_side	Inherited_from_father	Maternal_gene	Paternal_gene	E
	<int>	<chr>	<chr>	<chr>	<chr>	
6701	7	No	No	No	No	
6702	12	Yes	No	Yes	No	
6703	6	No	Yes	No	Yes	
6704	13	No	Yes	No	Yes	
6705	4	Yes	No	No	No	
6706	11	Yes	No	No	No	

In [11]:
```
#dimension of our df
dim(df)
```

6706 · 32

In [12]:
```
#Genetic Disorder depends upon disorder_subclass
ggplot(data = df, aes(Genetic_Disorder,fill = Disorder_Subclass)) +
geom_bar(width = 0.6)+  scale_x_discrete(guide = guide_axis(n.dodge = 3))
```



In [13]:
```
df <- df[, !names(df) %in% 'Genetic_Disorder']
```

In [14]:
```
head(df)
```

	Patient_Age	Genes_in_mother_s_side	Inherited_from_father	Maternal_gene	Paternal_gene	Bloo
	<int>	<chr>	<chr>	<chr>	<chr>	
1	11	No	No	Yes	No	
2	4	No	Yes	Yes	Yes	
3	1	Yes	Yes	No	No	
4	6	Yes	No	Yes	No	
5	10	Yes	Yes	Yes	No	
6	6	No	Yes	Yes	Yes	

In [15]:
```
head(df[, c(11:23), drop=FALSE])
```

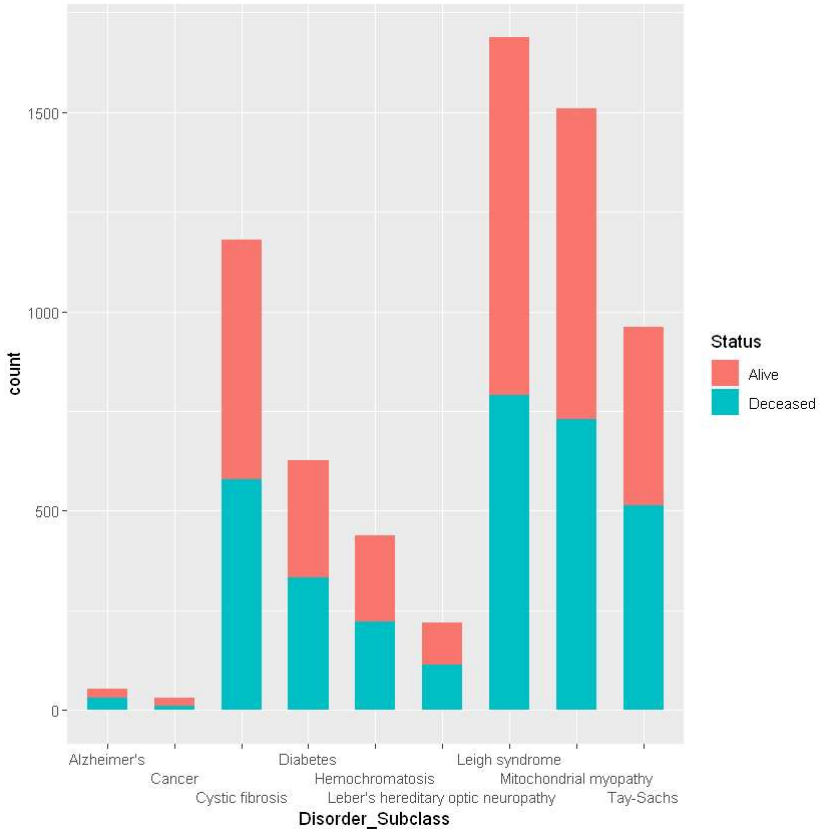
	Heart_Rate__rates_min	Follow.up	Gender	Birth_asphyxia	Autopsy_shows_birth_defect__if_app
	<chr>	<chr>	<chr>	<chr>	
1	Tachycardia	Low	Male	Yes	Not ap
2	Tachycardia	Low	Male	No	Not ap
3	Tachycardia	High	Male	Not available	
4	Tachycardia	High	Male	No	Not ap
5	Tachycardia	Low	Ambiguous	No	
6	Normal	Low	Ambiguous	No	

In [23]:
```
#based on domain information
drop <- c("H_O_radiation_exposure__x.ray_" ,"H_O_substance_abuse")
df = df[,!(names(df) %in% drop)]
```

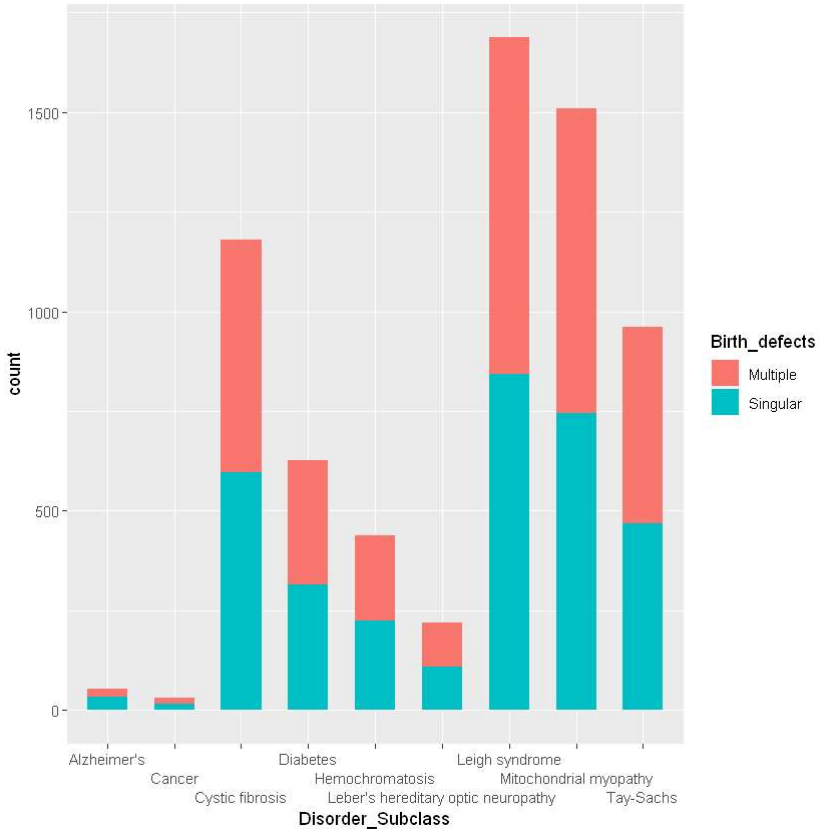
In [24]:
```
#correlation matrix for numeric columns
numeric_columns <- sapply(df, is.numeric)
cor_m <- cor(df[, numeric_columns])
cor_m
```

	Patient_Age	Blood_cell_count__mcL_	Mother_s_ag
Patient_Age	1.000000000	-0.006653421	-0.01488138
Blood_cell_count__mcL_	-0.006653421	1.000000000	0.00758383
Mother_s_age	-0.014881381	0.007583837	1.00000000
Father_s_age	-0.005919396	0.012839825	-0.00026434
No._of_previous_abortion	0.004241846	-0.005118477	0.00870984
White_Blood_cell_count__thousand_per_microliter_	-0.007781066	0.003946293	0.00280120
Symptom_1	0.013032928	0.017128502	-0.00281246
Symptom_2	0.011491371	-0.005550782	-0.00010648
Symptom_3	-0.015801886	-0.002605053	-0.00160556
Symptom_4	-0.006593085	0.011353858	-0.01901121
Symptom_5	-0.017132163	0.003202254	0.01310866

In [25]:
```
#visualization

ggplot(data = df, aes(Disorder_Subclass,fill =Status )) +
geom_bar(width = 0.6)+  scale_x_discrete(guide = guide_axis(n.dodge = 3))
```

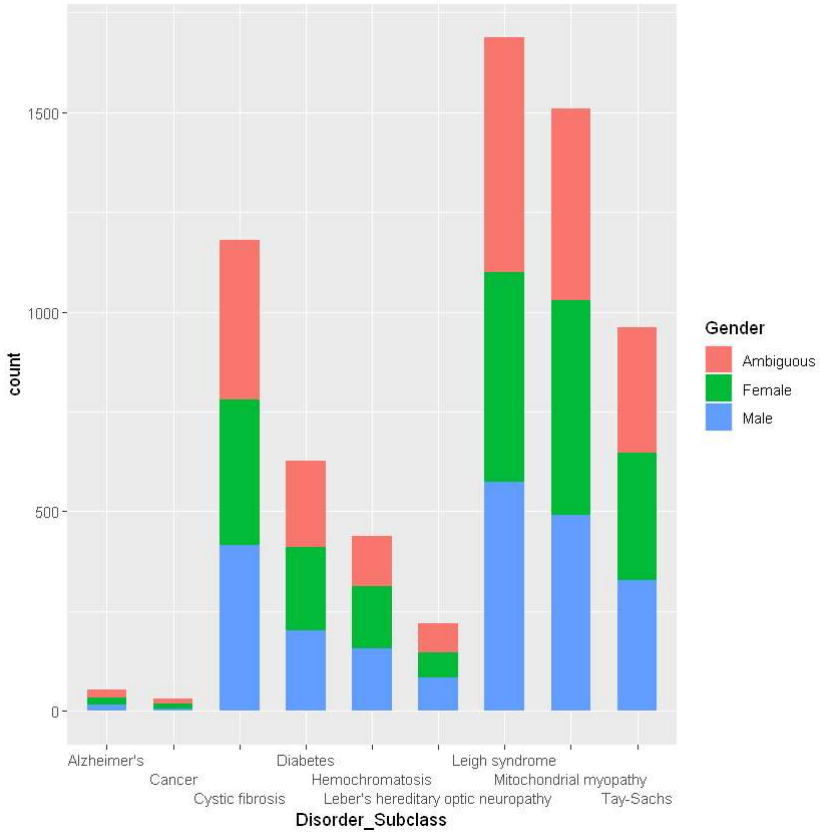


In [26]:
```
ggplot(data = df, aes(Disorder_Subclass,fill =Birth_defects )) +
geom_bar(width = 0.6)+  scale_x_discrete(guide = guide_axis(n.dodge = 3))
```

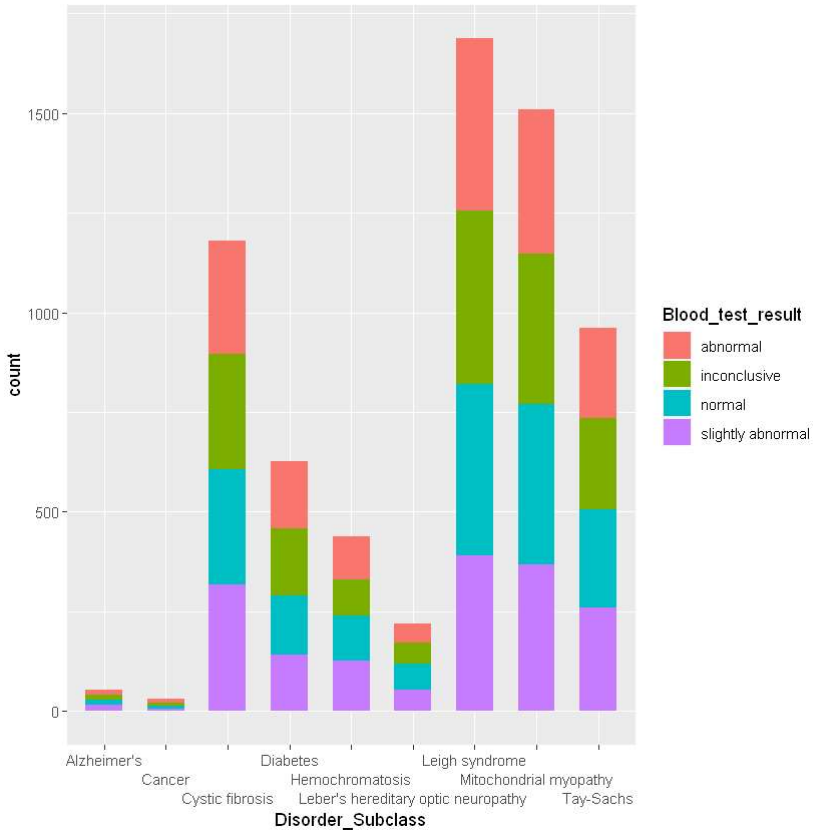


In [27]:
```
ggplot(data = df, aes(Disorder_Subclass,fill = Gender)) +
```

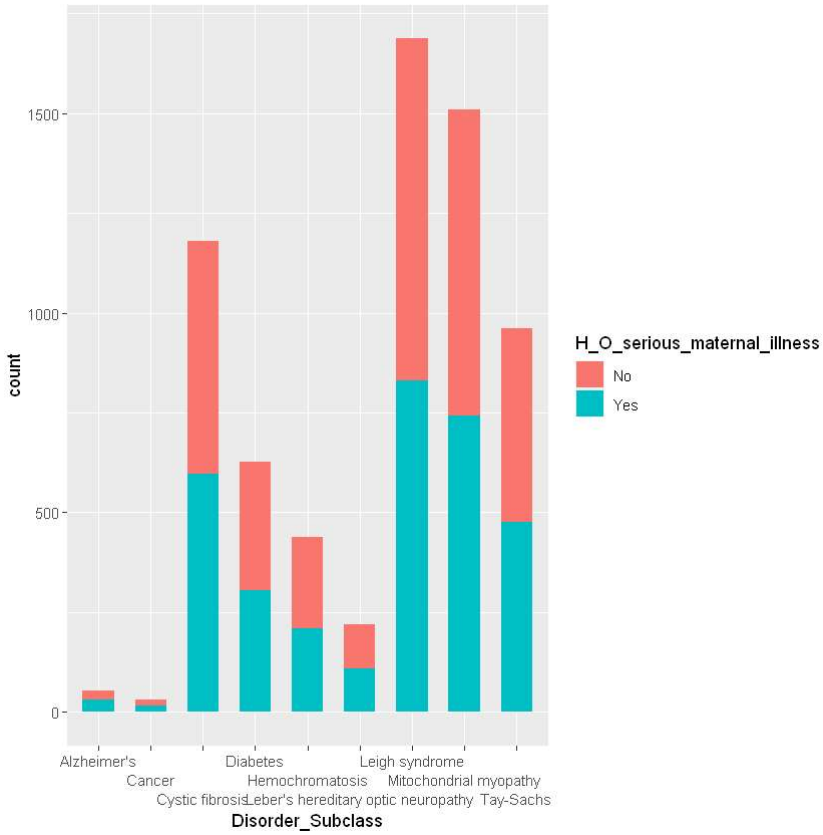
```
geom_bar(width = 0.6)+  scale_x_discrete(guide = guide_axis(n.dodge = 3))
```



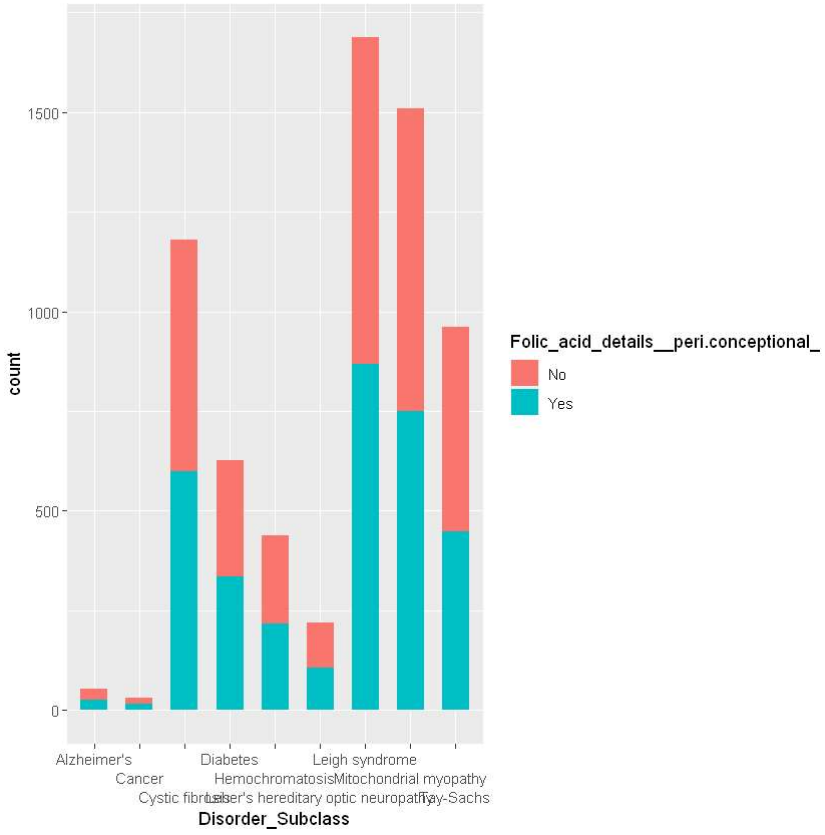
In [28]:
```
ggplot(data = df, aes( Disorder_Subclass ,fill =Blood_test_result )) +
geom_bar(width = 0.6)+  scale_x_discrete(guide = guide_axis(n.dodge = 3))
```



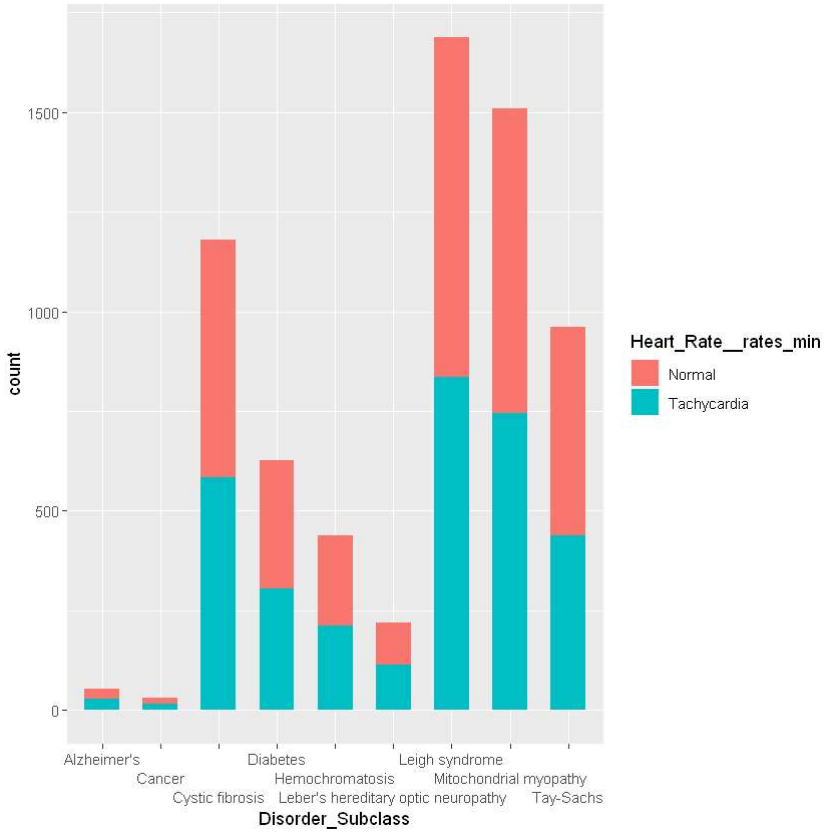
In [29]:
```
#df <- df[!apply(df, 1, function(row) any(row %in% c("not applicable", "NA"))), ]

ggplot(data = df, aes( Disorder_Subclass ,fill =H_O_serious_maternal_illness )) +
geom_bar(width = 0.6)+  scale_x_discrete(guide = guide_axis(n.dodge = 3))
```



In [30]:
```
ggplot(data = df, aes( Disorder_Subclass ,fill =Folic_acid_details__peri.conception
geom_bar(width = 0.6)+  scale_x_discrete(guide = guide_axis(n.dodge = 3))
```

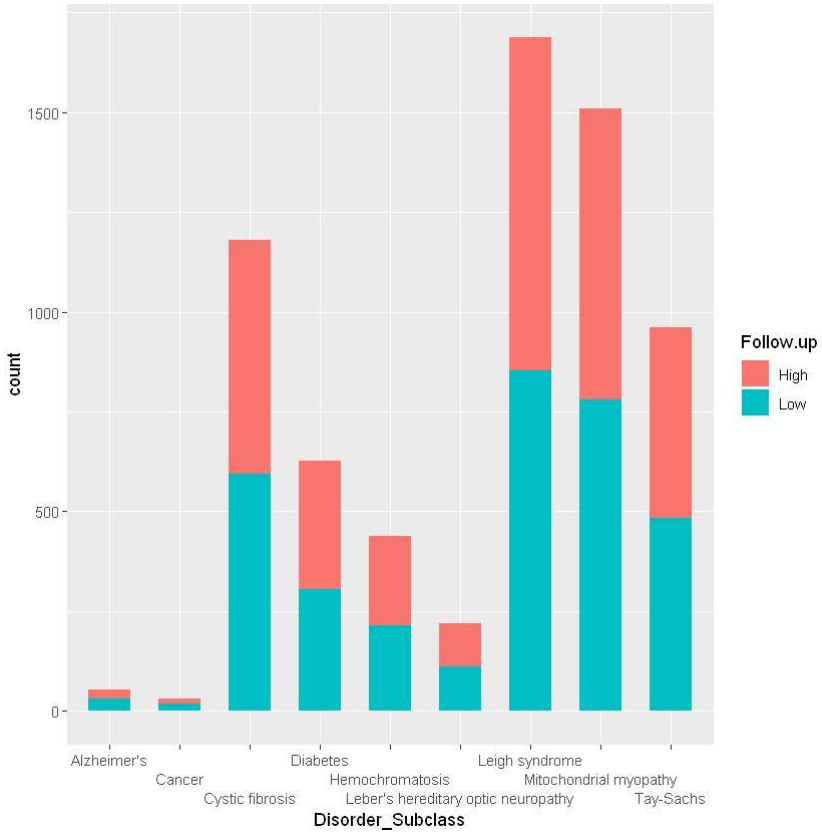


In [31]:
```
ggplot(data = df, aes( Disorder_Subclass ,fill = Heart_Rate__rates_min )) +
geom_bar(width = 0.6)+  scale_x_discrete(guide = guide_axis(n.dodge = 3))
```

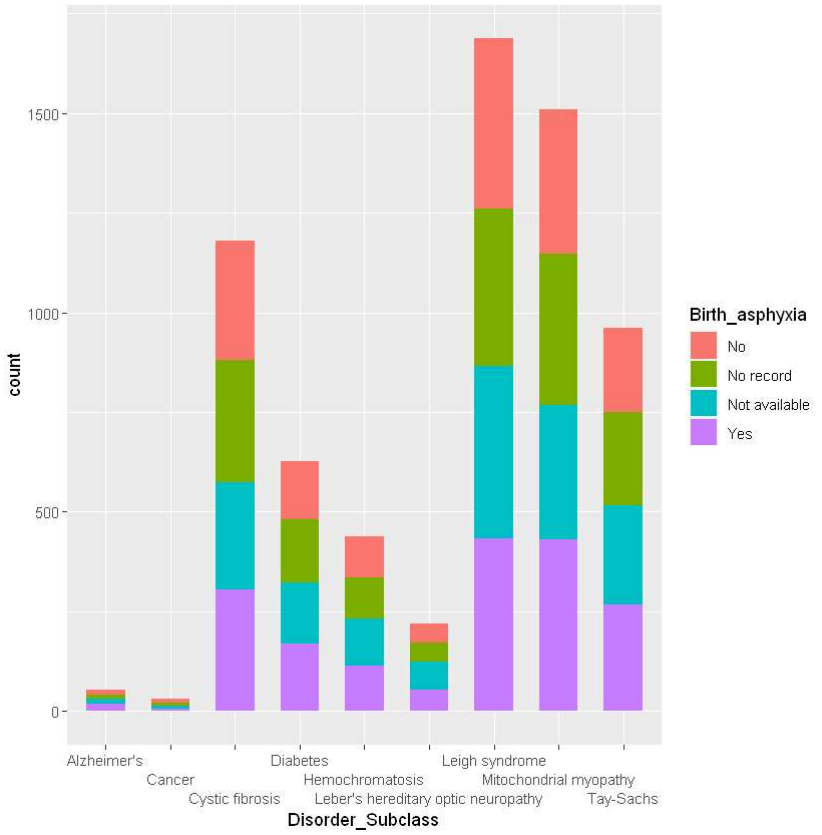


In [32]:
```
ggplot(data = df, aes( Disorder_Subclass ,fill = Follow.up )) +
```

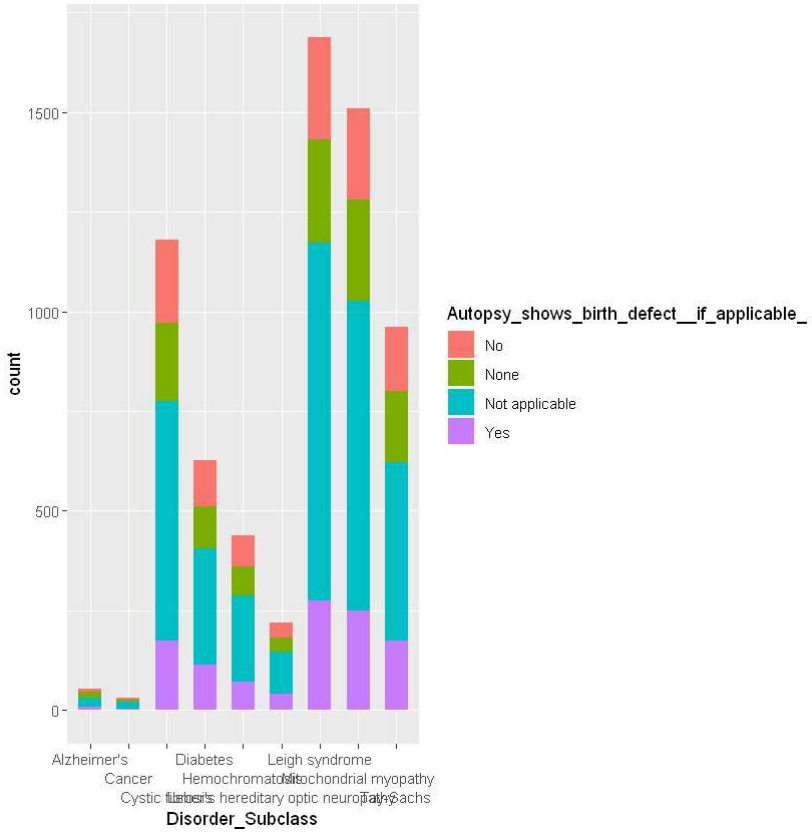
```
geom_bar(width = 0.6)+  scale_x_discrete(guide = guide_axis(n.dodge = 3))
```



In [33]:
```
ggplot(data = df, aes( Disorder_Subclass ,fill = Birth_asphyxia )) +
geom_bar(width = 0.6)+  scale_x_discrete(guide = guide_axis(n.dodge = 3))
```



In [34]:
```
ggplot(data = df, aes( Disorder_Subclass ,fill = Autopsy_shows_birth_defect__if_app
geom_bar(width = 0.6)+  scale_x_discrete(guide = guide_axis(n.dodge = 3))
```

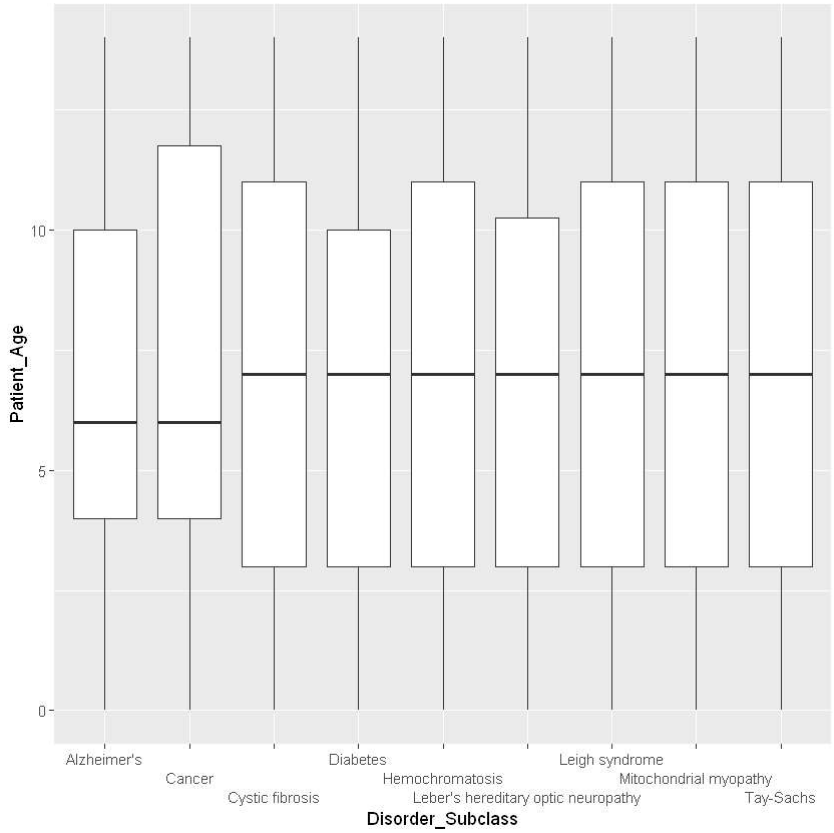


In [35]:
```
drop <- c("Birth_asphyxia" ,"Autopsy_shows_birth_defect__if_applicable_","Follow.up
          "Folic_acid_details__peri.conceptional_ ","H_O_serious_maternal_illness",
         "Blood_test_result","Gender")
df = df[,!(names(df) %in% drop)]
```

In [36]:
```
head(df)
```

	Patient_Age	Genes_in_mother_s_side	Inherited_from_father	Maternal_gene	Paternal_gene	Bloo
	<int>	<chr>	<chr>	<chr>	<chr>	
1	11	No	No	Yes	No	
2	4	No	Yes	Yes	Yes	
3	1	Yes	Yes	No	No	
4	6	Yes	No	Yes	No	
5	10	Yes	Yes	Yes	No	
6	6	No	Yes	Yes	Yes	

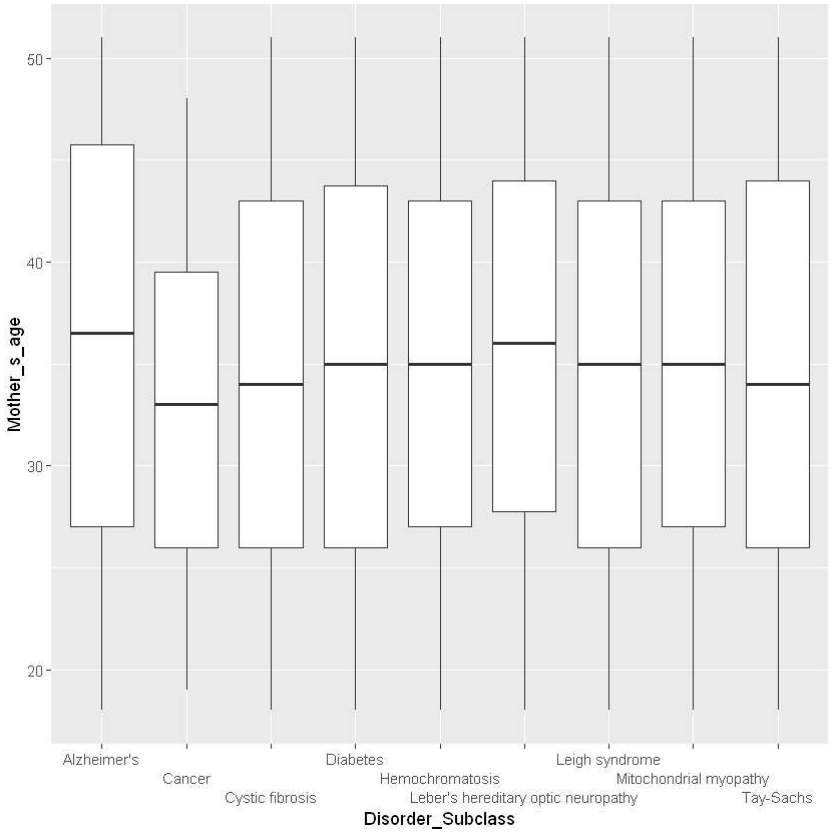
In [37]:
```
ggplot(df, aes(x = Disorder_Subclass, y = Patient_Age)) +
  geom_boxplot()+  scale_x_discrete(guide = guide_axis(n.dodge = 3))
```



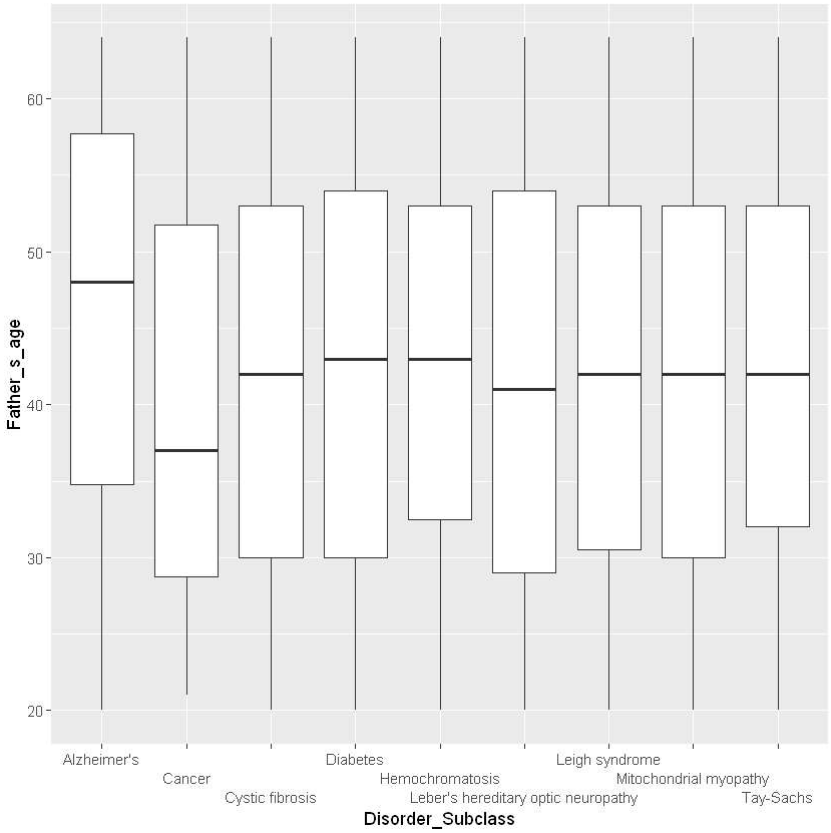
In [38]:
```
numeric_columns <- sapply(df, is.numeric)
cor(df[, numeric_columns])
```

	Patient_Age	Blood_cell_count__mcL_	Mother_s_ag
Patient_Age	1.000000000	-0.006653421	-0.01488138
Blood_cell_count__mcL_	-0.006653421	1.000000000	0.00758383
Mother_s_age	-0.014881381	0.007583837	1.00000000
Father_s_age	-0.005919396	0.012839825	-0.00026434
No._of_previous_abortion	0.004241846	-0.005118477	0.00870984
White_Blood_cell_count__thousand_per_microliter_	-0.007781066	0.003946293	0.00280120
Symptom_1	0.013032928	0.017128502	-0.00281246
Symptom_2	0.011491371	-0.005550782	-0.00010648
Symptom_3	-0.015801886	-0.002605053	-0.00160556
Symptom_4	-0.006593085	0.011353858	-0.01901121
Symptom_5	-0.017132163	0.003202254	0.01310866

In [39]:
```
ggplot(df, aes(x = Disorder_Subclass, y = Mother_s_age)) +
  geom_boxplot()+  scale_x_discrete(guide = guide_axis(n.dodge = 3))
```

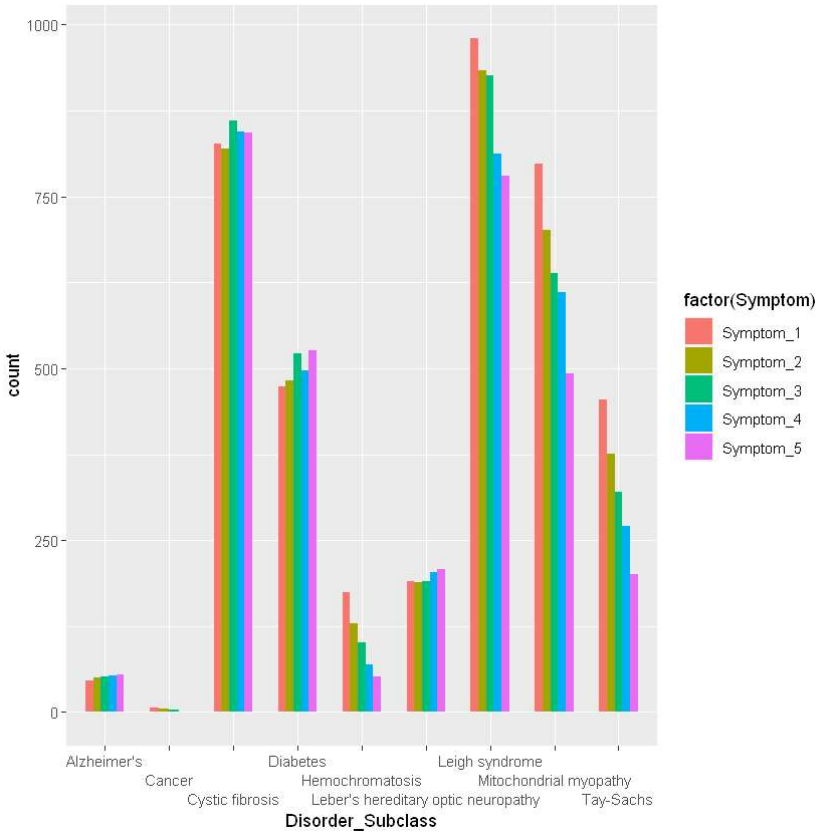


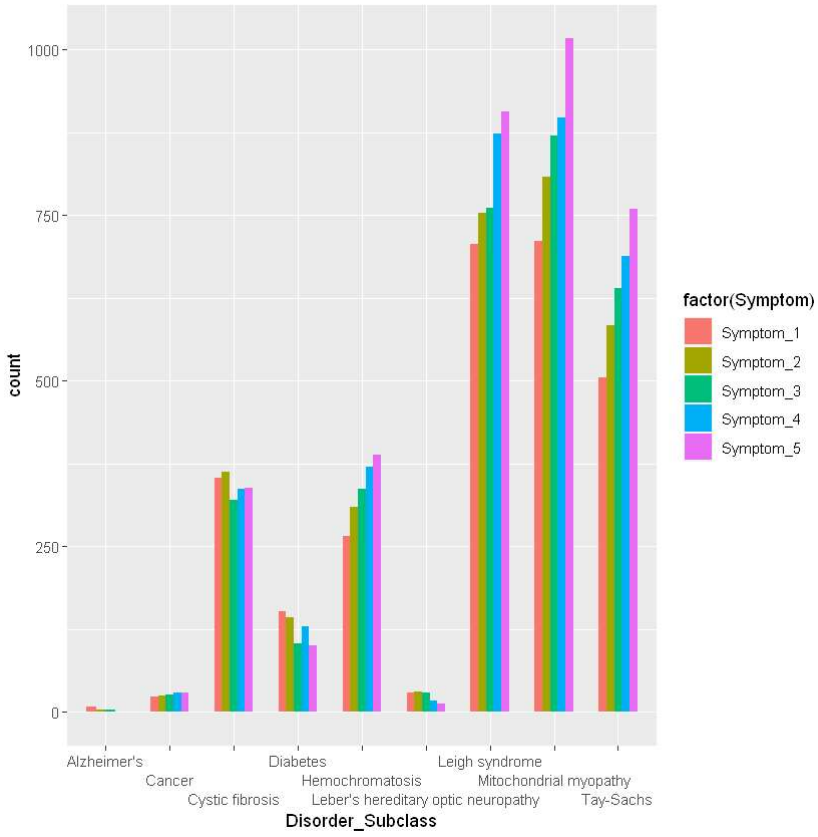
In [40]:
```
ggplot(df, aes(x = Disorder_Subclass, y = Father_s_age)) +
  geom_boxplot()+  scale_x_discrete(guide = guide_axis(n.dodge = 3))
```



In [41]:
```
df_long <- df %>%
  pivot_longer(cols = starts_with("Symptom"), names_to = "Symptom", values_to = "Va
```

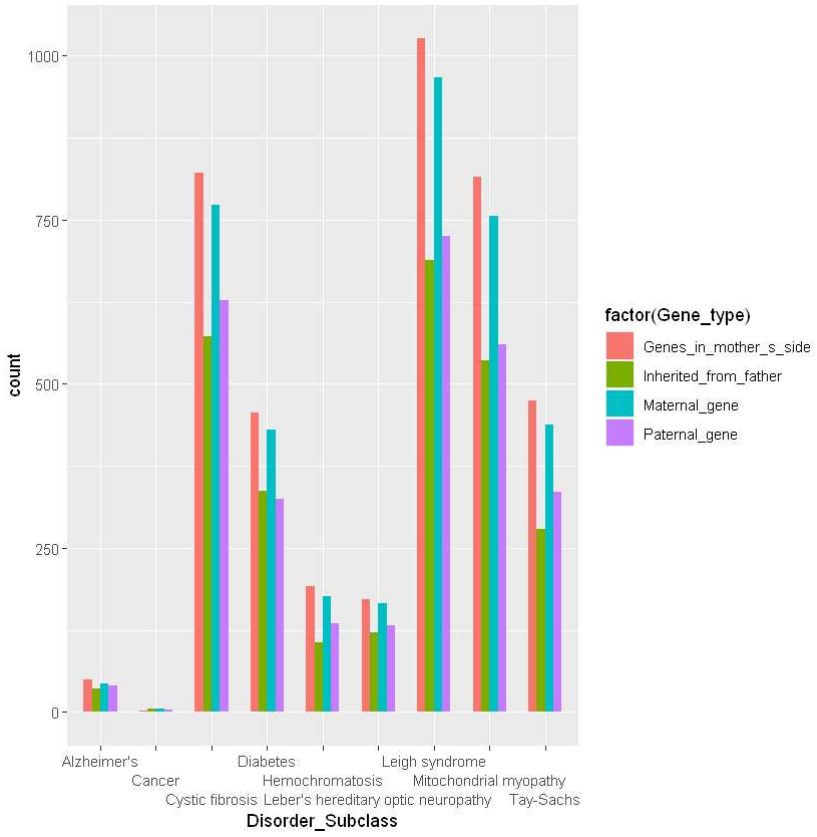
```r
df_filtered <- df_long %>%
  filter(Value == 1)

p1 <-ggplot(data = df_long, aes(Disorder_Subclass, fill = factor(Symptom), group =
  geom_bar(data = df_filtered, position = "dodge", width = 0.6, stat = "count") +
  scale_x_discrete(guide = guide_axis(n.dodge = 3))


df_filtered <- df_long %>%
  filter(Value == 0)

p2<-ggplot(data = df_long, aes(Disorder_Subclass, fill = factor(Symptom), group = S
  geom_bar(data = df_filtered, position = "dodge", width = 0.6, stat = "count") +
  scale_x_discrete(guide = guide_axis(n.dodge = 3))

p1
p2
```

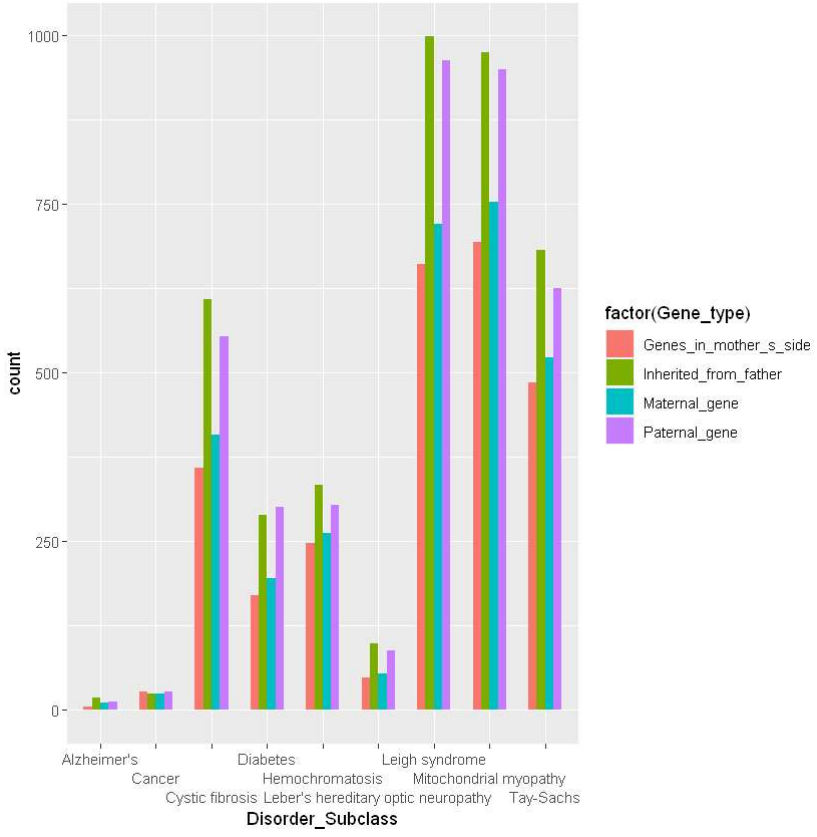




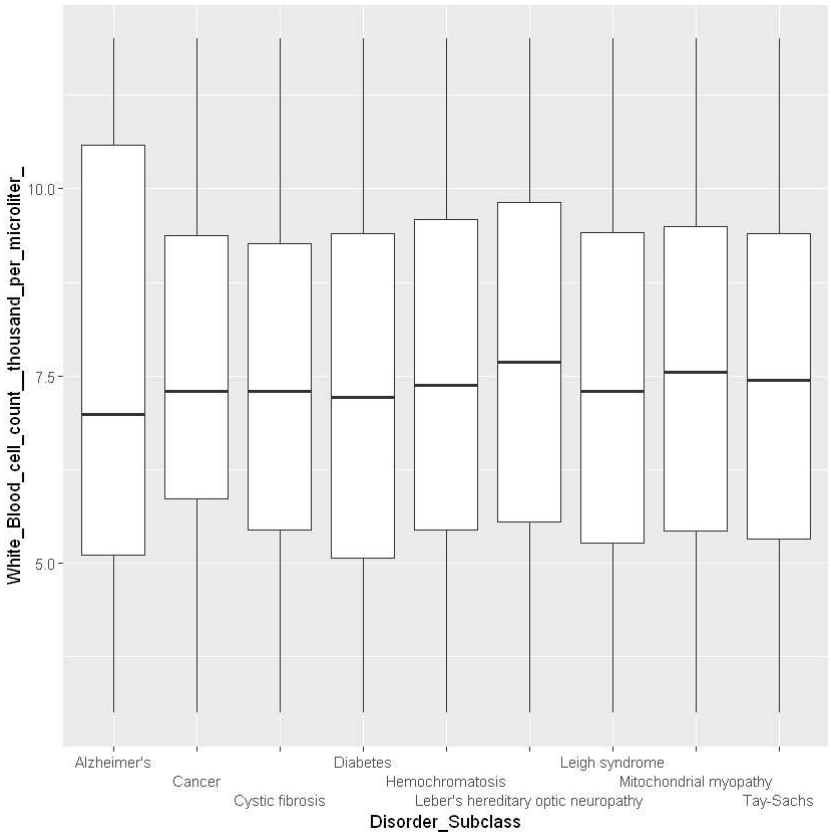
In [ ]:

In [42]:
```
df_long <- df %>%
  pivot_longer(cols = c(Genes_in_mother_s_side, Inherited_from_father, Maternal_gen
               names_to = "Gene_type", values_to = "Value")

df_long$Value <- factor(df_long$Value, levels = c("Yes", "No"))

df_filtered <- df_long %>%
  filter(Value == 'Yes')

p3<-ggplot(data = df_long, aes(Disorder_Subclass, fill = factor(Gene_type), group =
  geom_bar(data = df_filtered, position = "dodge", width = 0.6, stat = "count") +
  scale_x_discrete(guide = guide_axis(n.dodge = 3))

df_filtered <- df_long %>%
  filter(Value == 'No')

p4 <- ggplot(data = df_long, aes(Disorder_Subclass, fill = factor(Gene_type), group
  geom_bar(data = df_filtered, position = "dodge", width = 0.6, stat = "count") +
  scale_x_discrete(guide = guide_axis(n.dodge = 3))

p3
p4
```





In [43]:
```
ggplot(df, aes(x = Disorder_Subclass, y = White_Blood_cell_count__thousand_per_micr
  geom_boxplot()+  scale_x_discrete(guide = guide_axis(n.dodge = 3))
```


White_Blood_cell_count__thousand_per_microliter_
10.0
7.5
5.0
Alzheimer's
Cancer
Cystic fibrosis
Diabetes
Hemochromatosis
Leber's hereditary optic neuropathy
Leigh syndrome
Mitochondrial myopathy
Tay-Sachs
Disorder_Subclass

In [ ]:

In [ ]: