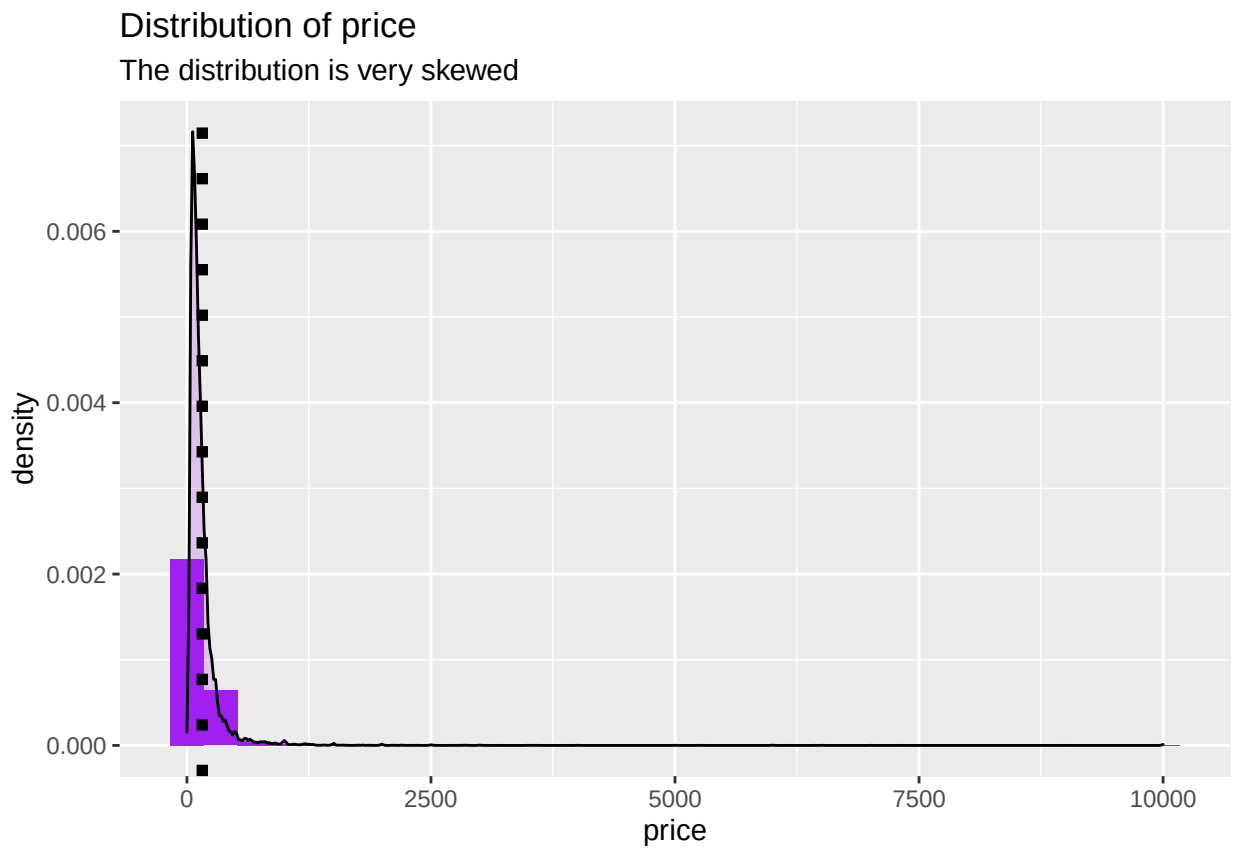
R Notebook- Inside_Airbnb_Data

```
data_NYC <- read.csv('listingsNY.csv')
summary(data_NYC)

glimpse(data_NYC)
dim(data_NYC)
summary(is.na(data_NYC))
```

```
unique(data_NYC$neighbourhood_group)
unique(data_NYC$neighbourhood)
```

```
ggplot(data_NYC, aes(price)) +
  geom_histogram(bins = 30, aes(y = ..density..), fill = "purple") +
  geom_density(alpha = 0.2, fill = "purple") +
  ggtitle("Distribution of price",
          subtitle = "The distribution is very skewed") +
  theme(axis.title = element_text(), axis.title.x = element_text()) +
  geom_vline(xintercept = round(mean(data_NYC$price), 2), size = 2, linetype = 3)
```



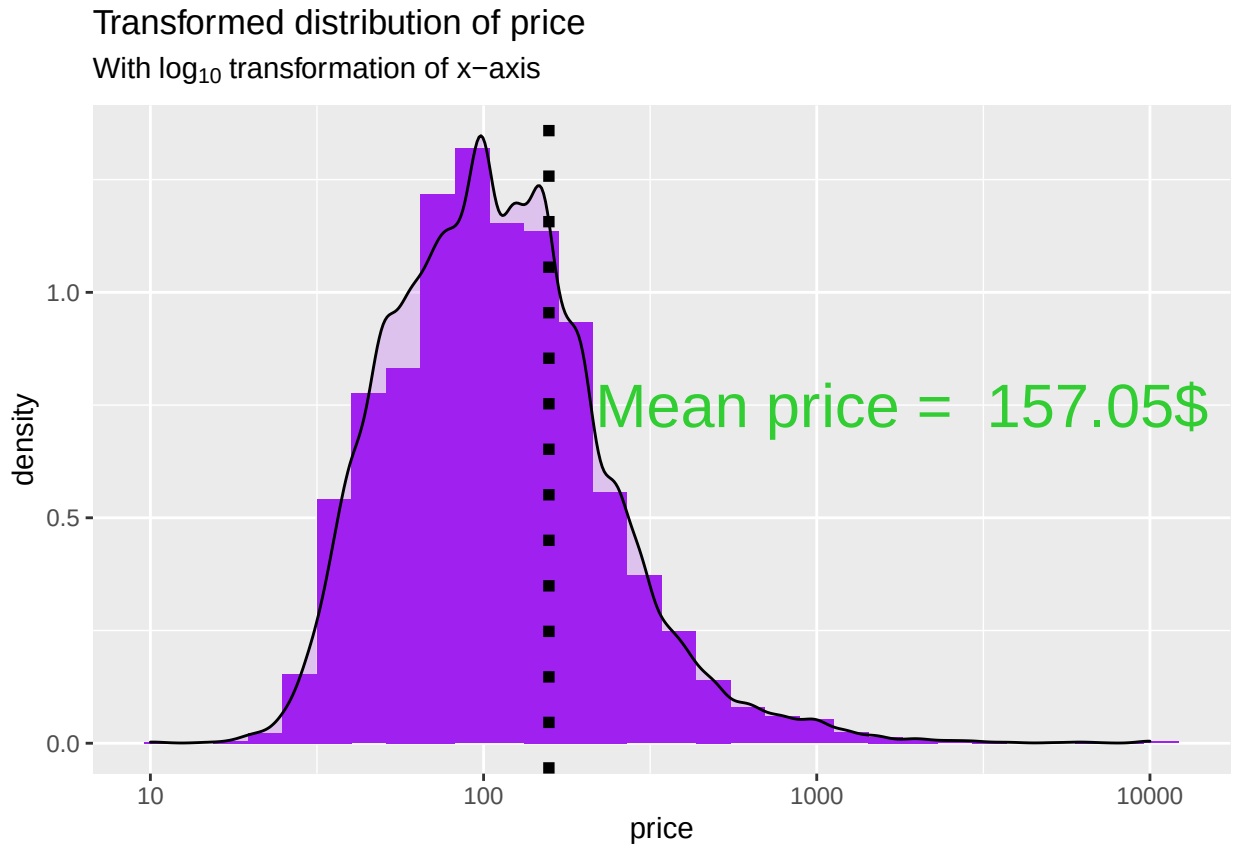
```
#Since the original distribution is very skewed, logarithmic transformation can be used to gain better
ggplot(data_NYC, aes(price)) +
  geom_histogram(bins = 30, aes(y = ..density..), fill = "purple") +
  geom_density(alpha = 0.2, fill = "purple") +
  ggtitle("Transformed distribution of price",
          subtitle = expression("With" ~'log'[10] ~ "transformation of x-axis")) +
  #theme(axis.title = element_text(), axis.title.x = element_text()) +
  geom_vline(xintercept = round(mean(data_NYC$price), 2), size = 2, linetype = 3) +
  scale_x_log10() +
  annotate("text", x = 1800, y = 0.75,label = paste("Mean price = ", paste0(round(mean(data_NYC$price),
           color =  "#32CD32", size = 8)
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## Warning: Removed 38 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 38 rows containing non-finite values (stat_density).
```




```
airbnb_by_neighbourhood <- data_NYC %>%
  group_by(neighbourhood_group) %>%
  summarise(price = round(mean(price), 2))


ggplot(data_NYC, aes(price)) +
  geom_histogram(bins = 30, aes(y = ..density..), fill = "purple") +
```

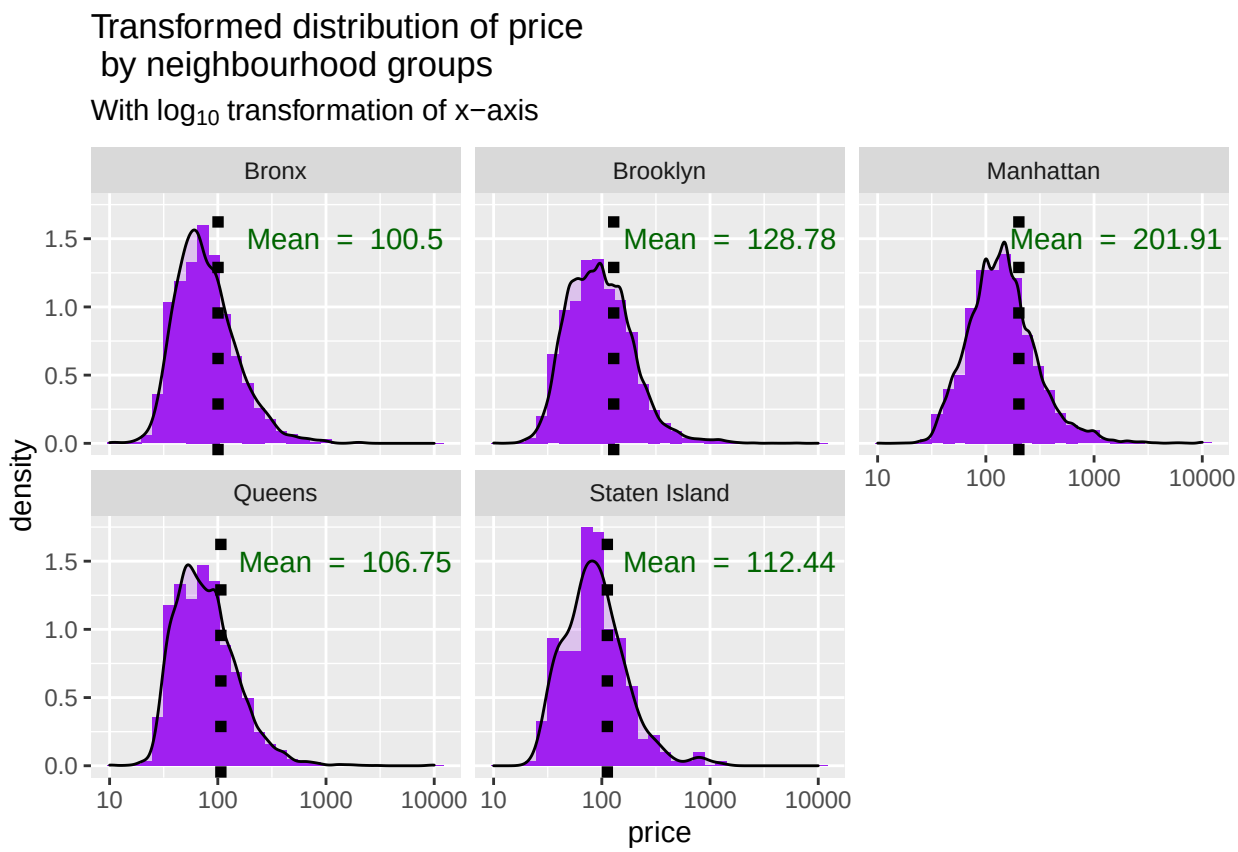
```
  geom_density(alpha = 0.2, fill = "purple") +
  ggtitle("Transformed distribution of price\n by neighbourhood groups",
          subtitle = expression("With" ~'log'[10] ~ "transformation of x-axis")) +
  geom_vline(data = airbnb_by_neighbourhood, aes(xintercept = price), size = 2, linetype = 3) +
  geom_text(data = airbnb_by_neighbourhood,y = 1.5, aes(x = price + 1400, label = paste("Mean  = ",price
  facet_wrap(~neighbourhood_group) +
  scale_x_log10()
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
```

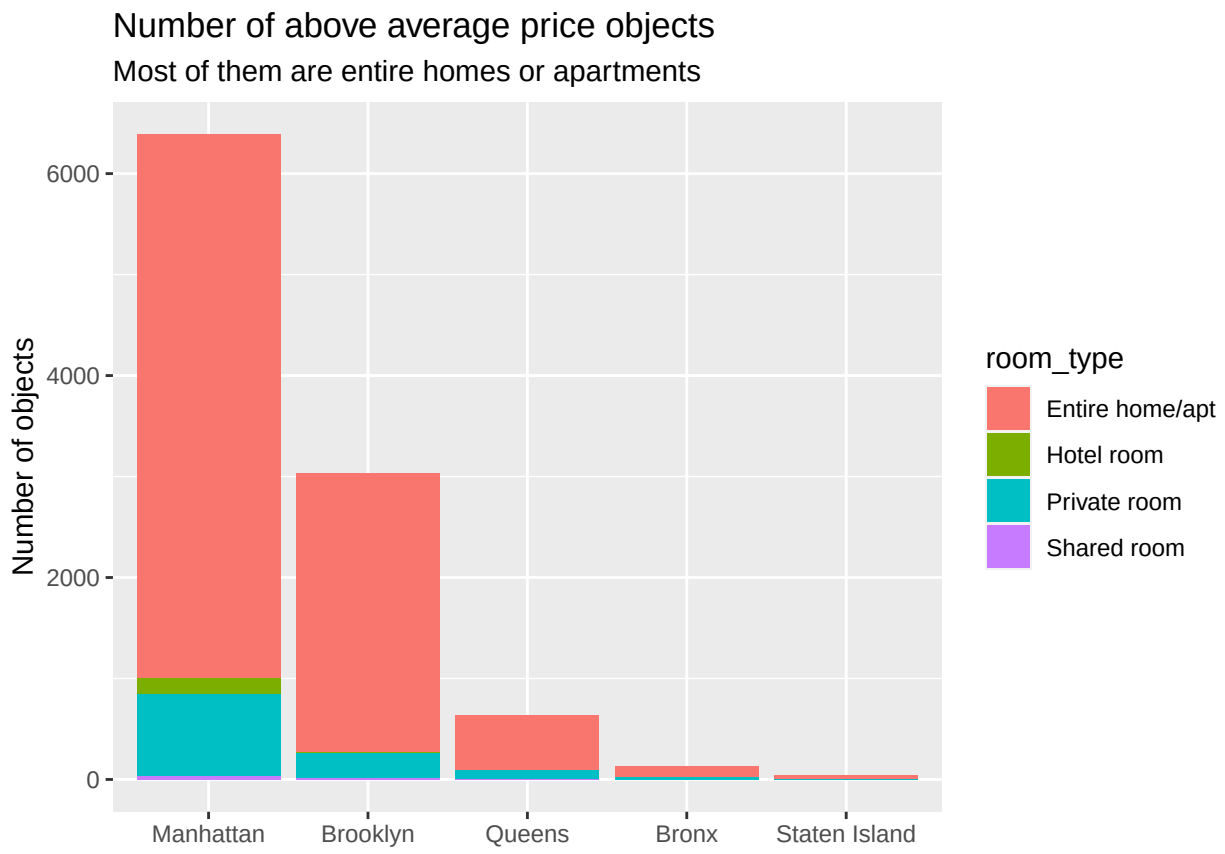
```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## Warning: Removed 38 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 38 rows containing non-finite values (stat_density).
```

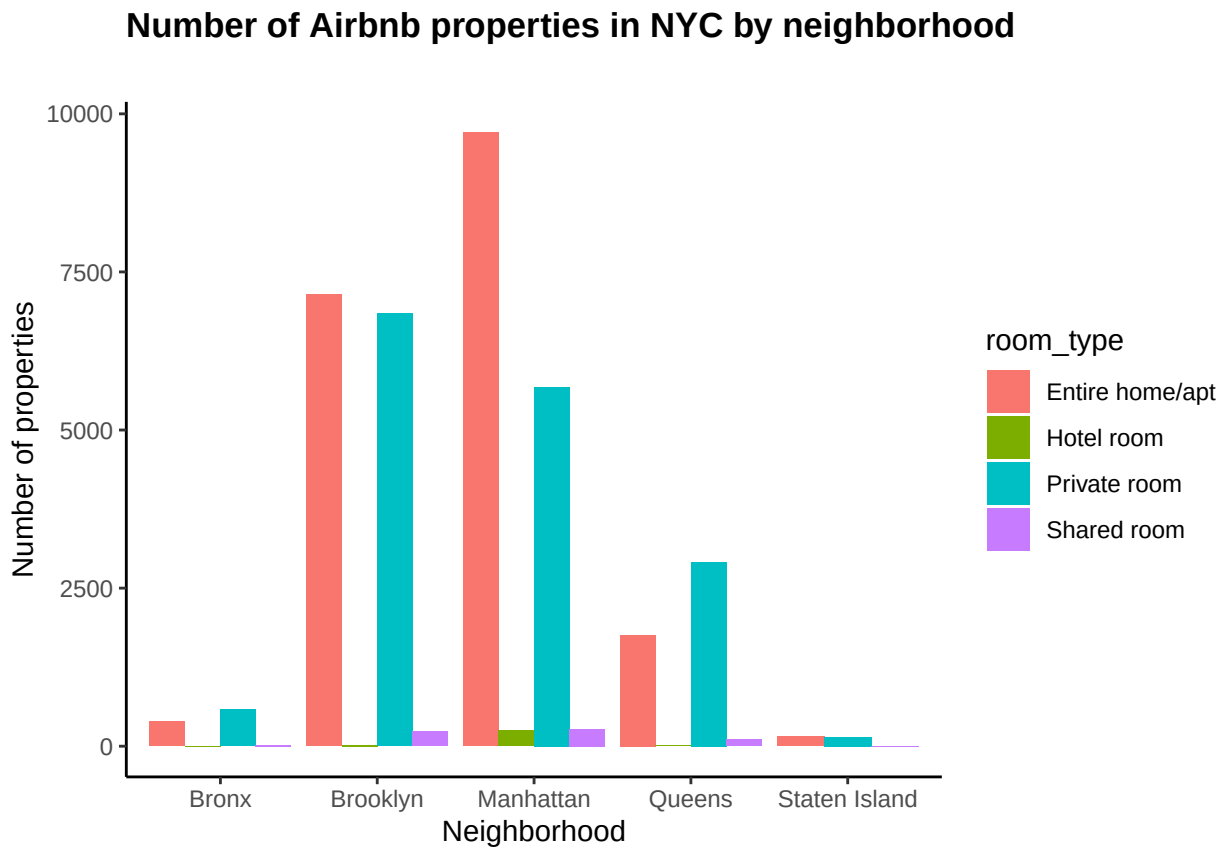



```
data_NYC %>% filter(price >= mean(price)) %>% group_by(neighbourhood_group, room_type) %>% tally %>%
  ggplot(aes(reorder(neighbourhood_group,desc(n)), n, fill = room_type)) +
  xlab(NULL) +
  ylab("Number of objects") +
  ggtitle("Number of above average price objects",
          subtitle = "Most of them are entire homes or apartments") +
  geom_bar(stat = "identity")
```

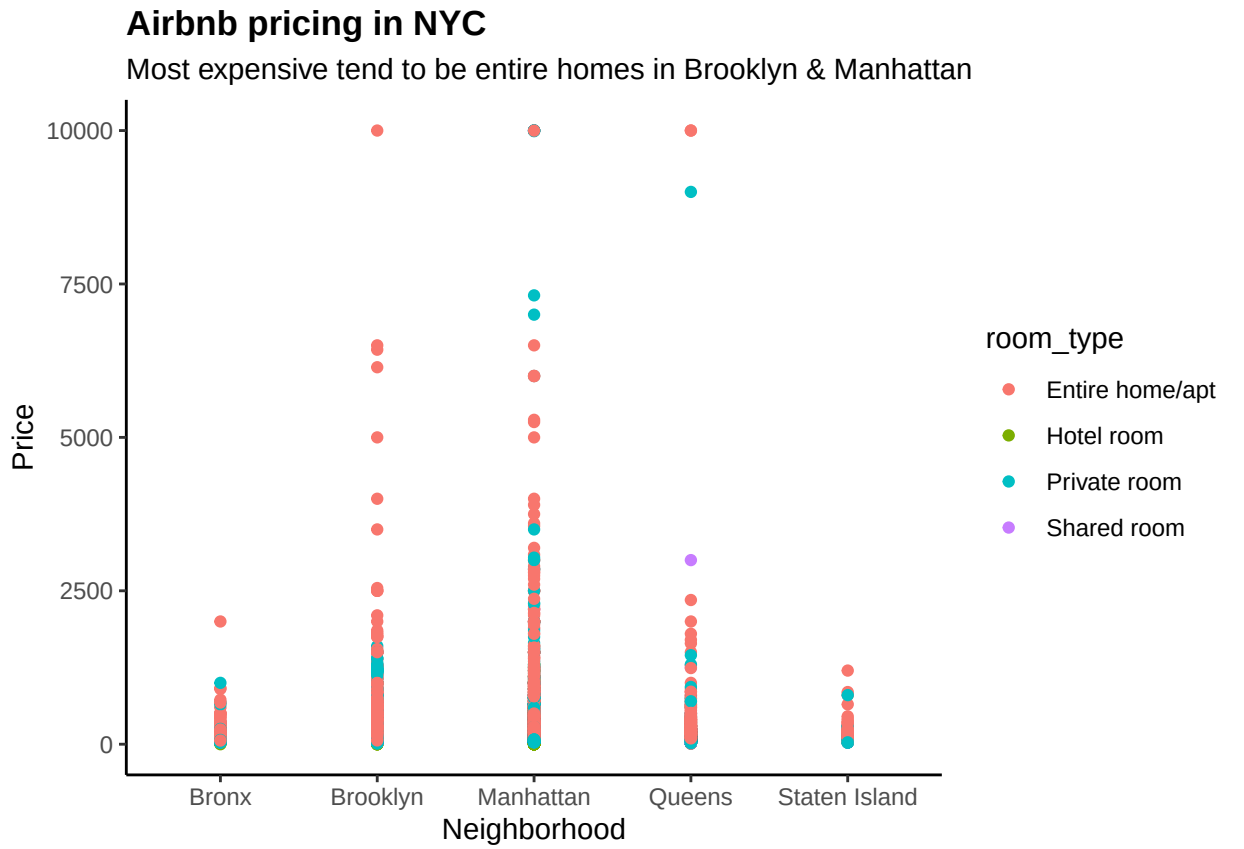



```
#
n_airbnb_by_area <- data_NYC %>%
  group_by(neighbourhood_group) %>%
  count(room_type)

ggplot(n_airbnb_by_area, aes(x = neighbourhood_group, y = n, fill = room_type)) +
  geom_bar(position = "dodge", stat = "identity") +
  theme_classic() +
  labs(title = "Number of Airbnb properties in NYC by neighborhood", subtitle = "", x = "Neighborhood",
  theme(plot.title = element_text(face = "bold"))
```

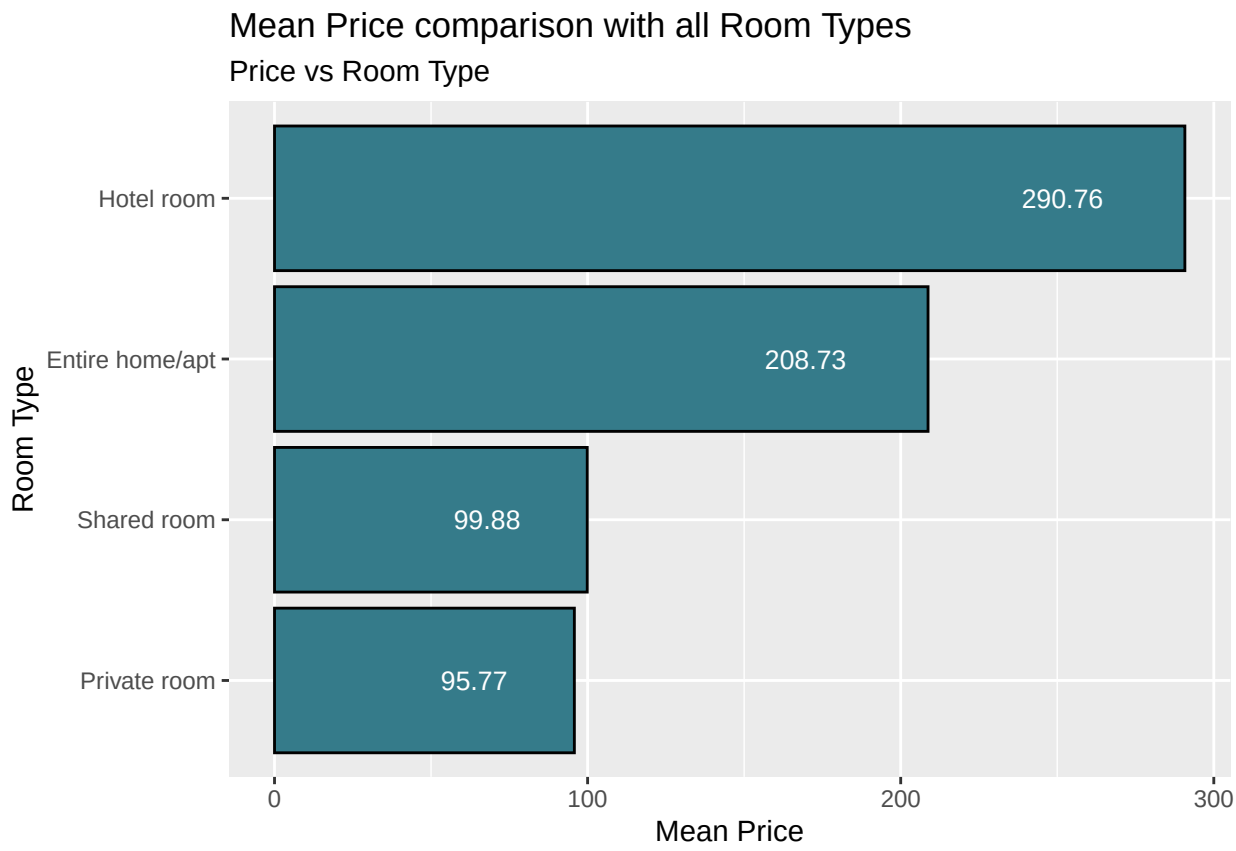



```
#
ggplot(data_NYC) +
  geom_point(aes(x = neighbourhood_group, y = price, color = room_type)) +
  theme_classic() +
  labs(title = "Airbnb pricing in NYC", subtitle = "Most expensive tend to be entire homes in Brooklyn &
  theme(plot.title = element_text(face = "bold"))
```




```
data_NYC %>%
  filter(!(is.na(room_type))) %>%
  filter(!(room_type == "Unknown")) %>%
  group_by(room_type) %>%
  summarise(mean_price = mean(price, na.rm = TRUE)) %>%
  ggplot(aes(x = reorder(room_type, mean_price), y = mean_price, fill = room_type)) +
  geom_col(stat ="identity", color = "black", fill="#357b8a") +
  coord_flip() +
  theme_gray() +
  labs(x = "Room Type", y = "Price") +
  geom_text(aes(label = round(mean_price,digit = 2)), hjust = 2.0, color = "white", size = 3.5) +
  ggtitle("Mean Price comparison with all Room Types", subtitle = "Price vs Room Type") +
  xlab("Room Type") +
  ylab("Mean Price")
```

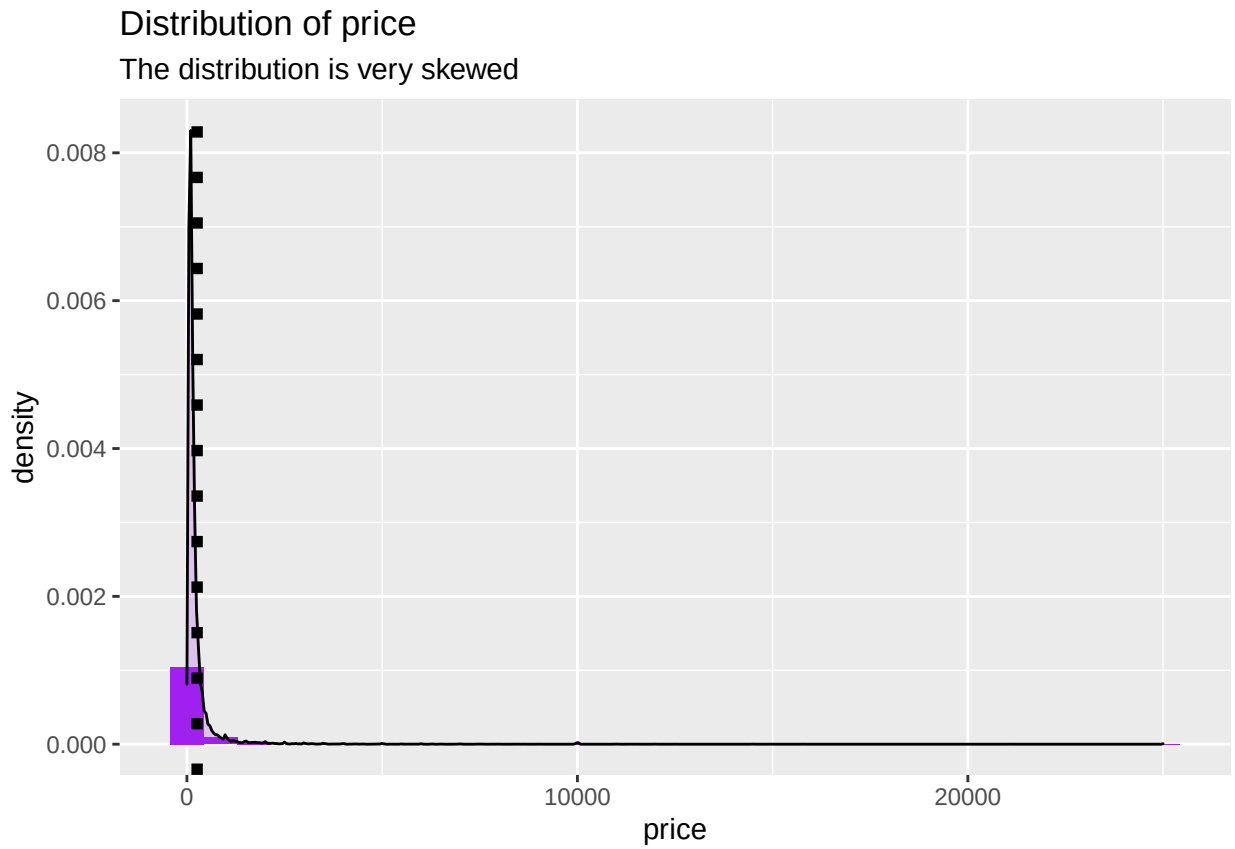
```
## Warning: Ignoring unknown parameters: stat
```




```
data_LA <- read.csv('listingsLA.csv')
summary(data_LA)

glimpse(data_LA)
dim(data_LA)
summary(is.na(data_LA))
```

```
ggplot(data_LA, aes(price)) +
  geom_histogram(bins = 30, aes(y = ..density..), fill = "purple") +
  geom_density(alpha = 0.2, fill = "purple") +
  ggtitle("Distribution of price",
          subtitle = "The distribution is very skewed") +
  theme(axis.title = element_text(), axis.title.x = element_text()) +
  geom_vline(xintercept = round(mean(data_LA$price), 2), size = 2, linetype = 3)
```



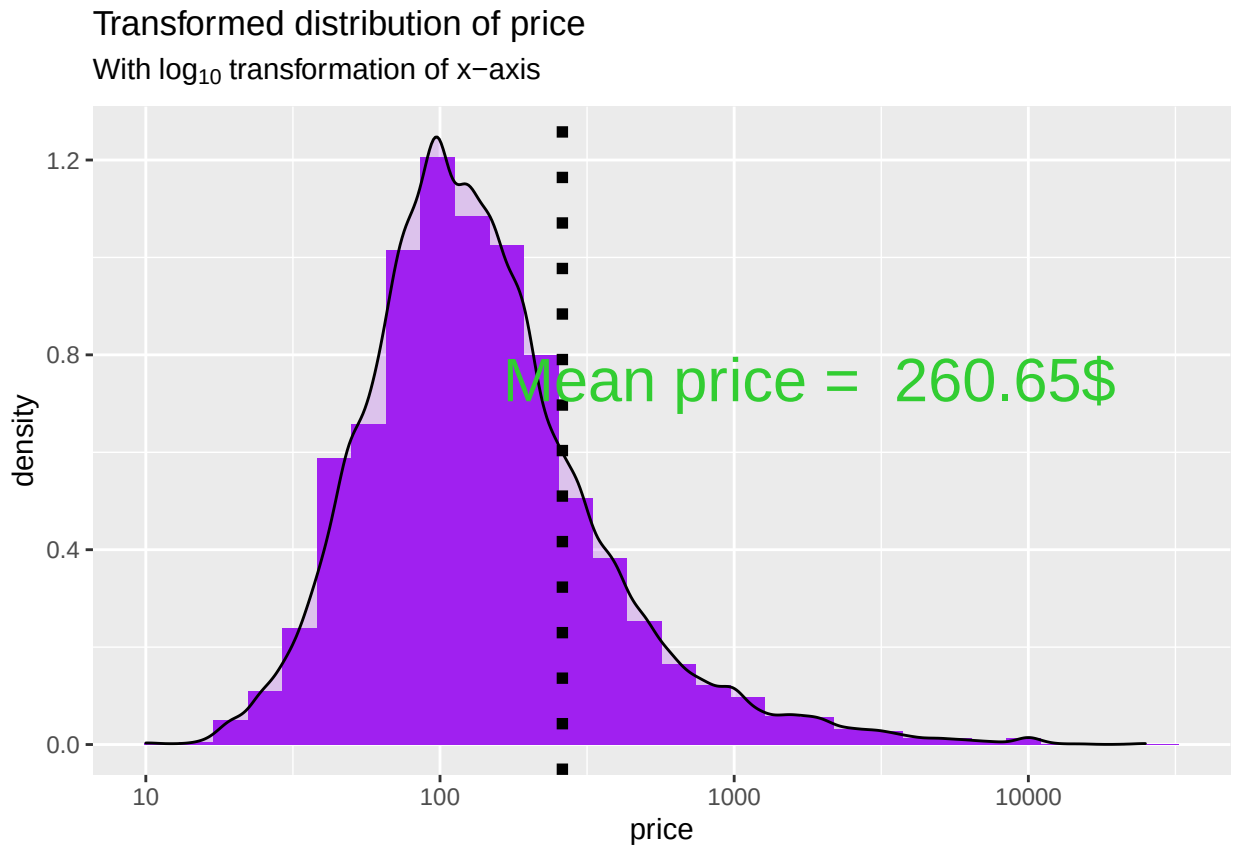

```
#Since the original distribution is very skewed, logarithmic transformation can be used to gain better
ggplot(data_LA, aes(price)) +
  geom_histogram(bins = 30, aes(y = ..density..), fill = "purple") +
  geom_density(alpha = 0.2, fill = "purple") +
  ggtitle("Transformed distribution of price",
          subtitle = expression("With" ~'log'[10] ~ "transformation of x-axis")) +
  #theme(axis.title = element_text(), axis.title.x = element_text()) +
  geom_vline(xintercept = round(mean(data_LA$price), 2), size = 2, linetype = 3) +
  scale_x_log10() +
  annotate("text", x = 1800, y = 0.75,label = paste("Mean price = ", paste0(round(mean(data_LA$price), 1
           color =  "#32CD32", size = 8)
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
```

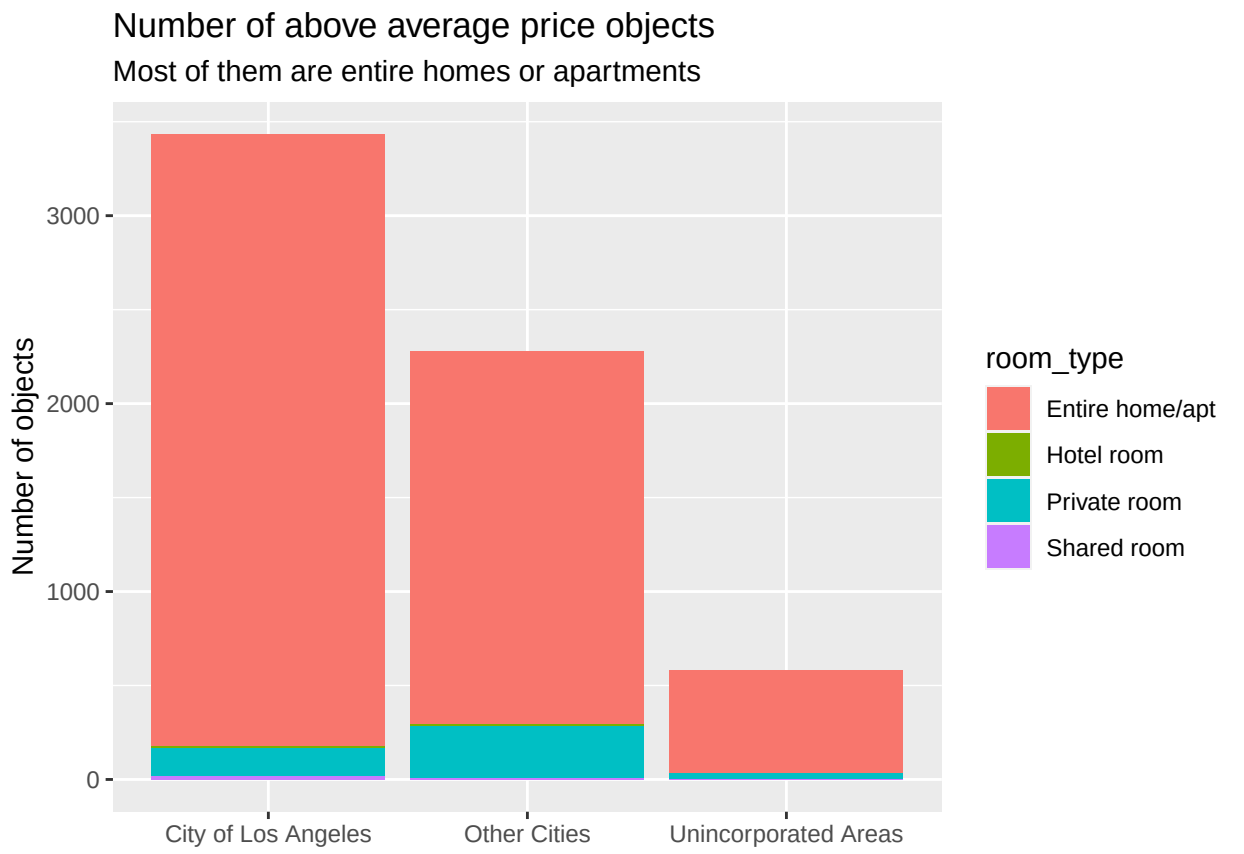
```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## Warning: Removed 16 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 16 rows containing non-finite values (stat_density).
```




```
data_LA %>% filter(price >= mean(price)) %>% group_by(neighbourhood_group, room_type) %>% tally %>%
  ggplot(aes(reorder(neighbourhood_group,desc(n)), n, fill = room_type)) +
  xlab(NULL) +
  ylab("Number of objects") +
  ggtitle("Number of above average price objects",
          subtitle = "Most of them are entire homes or apartments") +
  geom_bar(stat = "identity")
```




```
data_LA %>%
  group_by(neighbourhood_group) %>%
  summarize(min_price = min(price), max_price = max(price), avg_price = mean(price))
```

```
## # A tibble: 3 x 4
##   neighbourhood_group  min_price max_price avg_price
##   <chr>                    <int>     <int>     <dbl>
## 1 City of Los Angeles          0     21053      257.
## 2 Other Cities                 0     25000      272.
## 3 Unincorporated Areas        16     11335      235.
```

```
data_LA %>%
  filter(!(is.na(neighbourhood_group))) %>%
  filter(!(neighbourhood_group == "Unknown")) %>%
  group_by(neighbourhood_group) %>%
  summarise(mean_price = mean(price, na.rm = TRUE)) %>%
  ggplot(aes(x = reorder(neighbourhood_group, mean_price), y = mean_price, fill = neighbourhood_group))
  geom_col(stat ="identity", color = "black", fill="#357b8a") +
  coord_flip() +
  theme_gray() +
  labs(x = "Neighbourhood Group", y = "Price") +
  geom_text(aes(label = round(mean_price,digit = 2)), hjust = 2.0, color = "white", size = 3.5) +
  ggtitle("Mean Price comparison for each Neighbourhood Group - LA", subtitle = "Price vs Neighbourhood
  xlab("Neighbourhood Group") +
  ylab("Mean Price") +
  theme(legend.position = "none",
        plot.title = element_text(color = "black", size = 14, face = "bold", hjust = 0.5),
```

```
        plot.subtitle = element_text(color = "darkblue", hjust = 0.5),
        axis.title.y = element_text(),
        axis.title.x = element_text(),
        axis.ticks = element_blank())
```

```
## Warning: Ignoring unknown parameters: stat
```