

aqi

- 1. 问题
- 2. 导入包
- 3. 导入数据
- 4. 数据清洗
- 5. 探索性数据分析
 - 5.1 查看描述统计量
 - 5.2 单因子探索性数据分析
 - 1. 分析城市变量
 - 2. 分析地区变量
 - 3. 分析城市AQI变量
 - 4. 分析PM2.5变量
 - 5. 分析PM10变量
 - 6. 分析首要污染物变量
 - 7. 分析污染等级变量
 - 5.2 多因子探索性数据分析
- 参考资料

1. 问题

对该数据集，有以下问题需要解答：

1. 该数据集一共收集了多少个城市的空气数据？
2. 哪些城市的空气质量最好，哪个最差？
3. 北上广深这些一线城市的空气质量如何？
4. 城市AQI、PM2.5、PM10和地区AQI的数据分布是怎样的？
5. 污染等级中哪个级别最高？
6. PM2.5和PM10之间存在相关性吗？

2. 导入包

```
library(tidyverse) # 数据分析包
library(readxl) # 读取excel文件
library(psych) # 查看描述统计量
library(Hmisc)
library(pastecs)
library(knitr)
library(magrittr)
```

3. 导入数据

```
aqi <- read_xlsx("空气质量指数.xlsx")
head(aqi)
```

城市 <chr>	地区 <chr>	城市AQI <chr>	PM2.5浓度 <chr>	PM10浓度 <chr>	首要污染物 <chr>	污染等级 <chr>	地区AQI <dbl>
鞍山实时空气质量指数	明达新区	177	125µg/m³	228µg/m³	PM2.5	轻度污染	165
鞍山实时空气质量指数	千山	177	117µg/m³	145µg/m³	PM2.5	轻度污染	153
鞍山实时空气质量指数	深沟寺	177	138µg/m³	244µg/m³	PM2.5	轻度污染	183
鞍山实时空气质量指数	太平	177	126µg/m³	239µg/m³	PM2.5	轻度污染	166
鞍山实时空气质量指数	太阳城	177	142µg/m³	242µg/m³	PM2.5	轻度污染	189
鞍山实时空气质量指数	铁西工业园区	177	156µg/m³	324µg/m³	PM2.5	中度污染	206

6 rows

4. 数据清洗

- 将变量中的冗余字符去掉

```
aqi$城市 <- str_replace_all(aqi$城市, "[实时空气质量指数]", "") # 只保留城市名
aqi$PM2.5浓度 <- str_replace_all(aqi$PM2.5浓度, "[μg/m³|—μg/m³]", "") # 将单位去掉, 只保留数值
aqi$PM10浓度 <- str_replace_all(aqi$PM10浓度, "[μg/m³|—μg/m³]", "")
aqi$首要污染物 <- str_replace_all(aqi$首要污染物, "[—]", "NA")
```

- 转换变量的数据类型，以便进行更好地进行计算

```
aqi$城市AQI <- parse_double(aqi$城市AQI, na = "NA") # 转换为浮点类型

aqi$PM2.5浓度 <- parse_number(aqi$PM2.5浓度, na = "NA") # 转换为数值类型
aqi$PM10浓度 <- parse_number(aqi$PM10浓度, na = "NA")

aqi$首要污染物 <- parse_factor(aqi$首要污染物, na = "NA") # 转换为因子
level <- c("优", "良", "轻度污染", "中度污染", "严重污染")
aqi$污染等级 <- parse_factor(aqi$污染等级, levels = level, na = "NA")
```

- 简化变量名

```
aqi <- rename(aqi, PM2.5 = PM2.5浓度, PM10 = PM10浓度)
```

```
# 输出头6行数据
head(aqi)
```

城市 <chr>	地区 <chr>	城市AQI <dbl>	PM2.5 <dbl>	PM10 <dbl>	首要污染物 <fctr>	污染等级 <fctr>	地区AQI <dbl>
鞍山	明达新区	177	125	228	PM2.5	轻度污染	165
鞍山	千山	177	117	145	PM2.5	轻度污染	153
鞍山	深沟寺	177	138	244	PM2.5	轻度污染	183
鞍山	太平	177	126	239	PM2.5	轻度污染	166
鞍山	太阳城	177	142	242	PM2.5	轻度污染	189
鞍山	铁西工业园区	177	156	324	PM2.5	中度污染	206

6 rows

5. 探索性数据分析

5.1 查看描述统计量

```
stat.desc(aqi)
```

	城 市 <lgl>	地 区 <lgl>	城市AQI <dbl>	PM2.5 <dbl>	PM10 <dbl>	首要污 染物 <lgl>	污染 等级 <lgl>	地区AQI <dbl>
nbr.val	NA	NA	1.453000e+03	1.453000e+03	1.389000e+03	NA	NA	1.453000e+03
nbr.null	NA	NA	0.000000e+00	0.000000e+00	0.000000e+00	NA	NA	0.000000e+00
nbr.na	NA	NA	0.000000e+00	0.000000e+00	6.400000e+01	NA	NA	0.000000e+00
min	NA	NA	2.600000e+01	1.000000e+00	1.000000e+00	NA	NA	1.100000e+01
max	NA	NA	5.000000e+02	4.760000e+02	1.135000e+03	NA	NA	5.000000e+02
range	NA	NA	4.740000e+02	4.750000e+02	1.134000e+03	NA	NA	4.890000e+02
sum	NA	NA	1.242730e+05	8.403500e+04	1.360510e+05	NA	NA	1.248190e+05
median	NA	NA	7.600000e+01	5.000000e+01	8.700000e+01	NA	NA	7.500000e+01
mean	NA	NA	8.552856e+01	5.783551e+01	9.794888e+01	NA	NA	8.590434e+01
SE.mean	NA	NA	1.119277e+00	9.671441e-01	1.829588e+00	NA	NA	1.186049e+00

1-10 of 14 rows

Previous 1 2 Next

整个数据集有1453个观测（行），8个变量（列），这些变量分别是城市、地区、城市AQI、PM2.5、PM10、首要污染物、污染等级和地区AQI。

5.2 单因子探索性数据分析

1. 分析城市变量

- 统计城市数量

```
aqi %>%
  group_by(城市) %>%
  count() %>%
  summary()
```

```
##      城市              n
## Length:365      Min.   : 1.000
## Class :character 1st Qu.: 2.000
## Mode :character  Median : 4.000
##                Mean    : 3.981
##                3rd Qu.: 5.000
##                Max.    :17.000
```

经聚合后，数据集有365个城市。

2. 分析地区变量

- 统计地区数量

```
aqi %>%
  group_by(地区) %>%
  count() %>%
  summary()
```

```
##      地区              n
## Length:1264      Min.   : 1.00
## Class :character 1st Qu.: 1.00
## Mode :character  Median : 1.00
##                Mean    : 1.15
##                3rd Qu.: 1.00
##                Max.    :27.00
```

经聚合后，收集空气数据的地区有1264个。整个数据集有1453个观测，有些城市的数据收集地区有重复，重复数量有189个

3. 分析城市AQI变量

- 查看城市AQI的描述统计量

```
describe(aqi$城市AQI)
```

```
## aqi$城市AQI
##      n missing distinct      Info      Mean      Gmd      .05      .10
##    1453      0      129        1    85.53    42.64    39.0    44.0
##      .25      .50      .75      .90      .95
##    56.0     76.0    107.0    138.0    159.4
##
## lowest : 26 27 28 29 30, highest: 178 189 196 227 500
```

城市AQI变量中有1453个值，没有缺失值，其中平均值是85.53，中位数是76，最小值是26，最大值是500

- 按城市分组，计算各城市的AQI平均值

```
avg_city_aqi <- aqi %>%
  group_by(城市) %>%
  summarise(城市AQI平均值 = mean(城市AQI))
avg_city_aqi
```

城市 <chr>	城市AQI平均值 <dbl>
阿坝州	29
阿克苏地区	500
阿拉善盟	42
阿勒泰地区	49
阿里地区	28
安康	88
安庆	150

城市 <chr>	城市AQI平均值 <dbl>
安顺	41
安阳	100
鞍山	177
1-10 of 365 rows	
Previous	1 2 3 4 5 6 ... 37 Next

- 查看城市AQI平均值的描述统计量

```
summary(avg_city_aqi)
```

```
##      城市      城市AQI平均值
## Length:365      Min.   : 26.00
## Class :character 1st Qu.: 55.00
## Mode  :character Median : 74.00
##                Mean   : 83.19
##                3rd Qu.:102.00
##                Max.   :500.00
```

365个城市的AQI指数平均值为83.19，中位数是74，最小值是26，最大值是500。

- 城市AQI平均值的数值分布

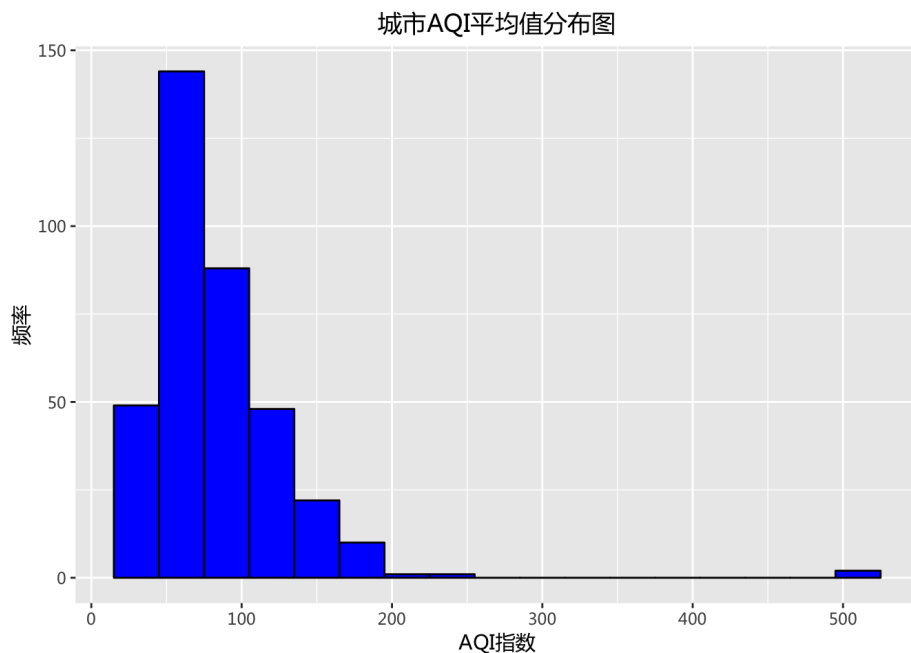
```
avg_city_aqi %>%
  count(AQI指数 = cut_width(城市AQI平均值, 30))
```

AQI指数 <fctr>	n <int>
[15,45]	49
(45,75]	144
(75,105]	88
(105,135]	48
(135,165]	22
(165,195]	10
(195,225]	1
(225,255]	1
(495,525]	2
9 rows	

- 绘制城市AQI平均值的直方图

```
plot_theme = theme(plot.title = element_text(hjust = 0.5),
  text = element_text(family = "MicrosoftYaHei"))

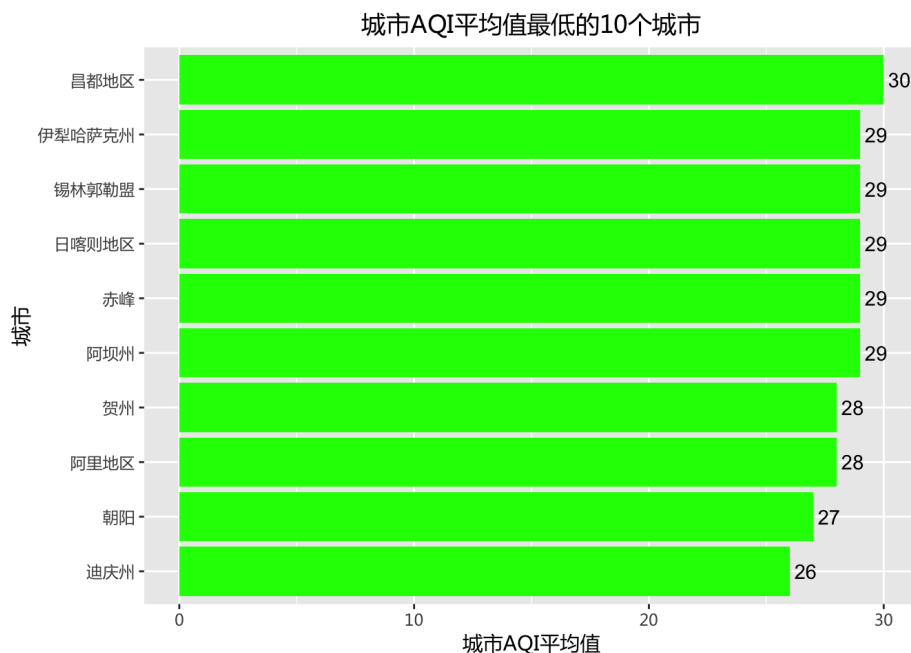
ggplot(avg_city_aqi, aes(城市AQI平均值)) +
  geom_histogram(color="black", fill="blue", binwidth = 30) +
  labs(title="城市AQI平均值分布图", x="AQI指数", y="频率") +
  plot_theme
```



有361个城市的AQI指数在15-195之间，占总体的98.9%，空气质量等级在优到轻度污染之间

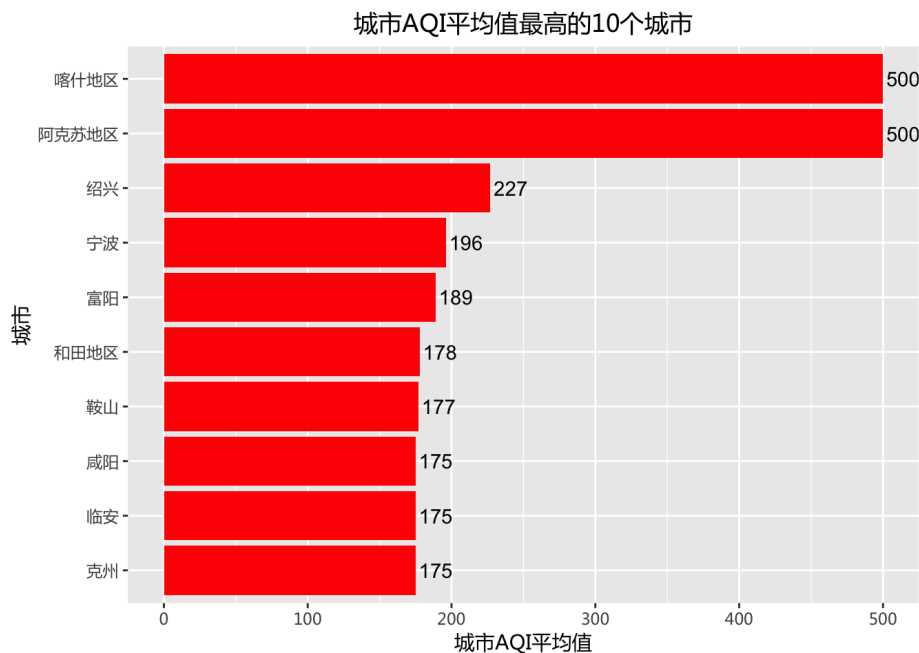
- 城市AQI平均值最低的10个城市

```
avg_city_aqi %>%  
  arrange(城市AQI平均值) %>% # 按从小到大排列  
  head(10) %>% # 输出AQI平均值最低的10个城市  
  # 绘制条形图  
  ggplot(aes(reorder(城市, 城市AQI平均值), 城市AQI平均值)) +  
  geom_bar(stat = "identity", fill = "green") +  
  labs(title = "城市AQI平均值最低的10个城市", x="城市", y="城市AQI平均值") + # 标题  
  geom_text(aes(label=城市AQI平均值), hjust=-0.2) + # 数据标签  
  coord_flip() + # 图形转置  
  plot_theme
```



- 城市AQI平均值最高的10个城市

```
avg_city_aqi %>%  
  arrange(desc(城市AQI平均值)) %>% # 按从大到小降序排列  
  head(10) %>% # 输出AQI平均值最高的10个城市  
  # 绘制条形图  
  ggplot(aes(reorder(城市, 城市AQI平均值), 城市AQI平均值)) +  
  geom_bar(stat = "identity", fill = "red") +  
  labs(title = "城市AQI平均值最高的10个城市", x="城市", y="城市AQI平均值") +  
  geom_text(aes(label=城市AQI平均值), hjust=-0.1) +  
  coord_flip() +  
  plot_theme
```



4. 分析PM2.5变量

- 查看PM2.5的描述统计量

```
describe(aqi$PM2.5)
```

```
## aqi$PM2.5
##      n missing distinct    Info      Mean      Gmd      .05      .10
##    1453      0      161      1    57.84    38.87     13     20
##      .25      .50      .75      .90      .95
##      32      50      79     107     125
##
## lowest :   1   2   3   4   5, highest: 193 212 272 283 476
```

PM2.5变量有1453个值，没有缺失值，其中平均值是57.84，中位数是50，最小值是1，最大值是476

- 按城市分组，计算各城市的PM2.5平均值

```
avg_city_pm2.5 <- aqi %>%
  group_by(城市) %>%
  summarise(城市PM2.5平均值 = mean(PM2.5))
avg_city_pm2.5
```

城市 <chr>	城市PM2.5平均值 <dbl>
阿坝州	2.666667
阿克苏地区	247.500000
阿拉善盟	23.000000
阿勒泰地区	16.500000
阿里地区	4.000000
安康	64.000000
安庆	114.500000
安顺	27.000000
安阳	75.200000
鞍山	134.000000

1-10 of 365 rows

Previous 1 2 3 4 5 6 ... 37 Next

- 查看城市PM2.5平均值的描述统计量

```
summary(avg_city_pm2.5)
```

```
##      城市      城市PM2.5平均值
## Length:365      Min.   :  2.667
## Class :character 1st Qu.: 30.500
## Mode  :character Median : 47.125
##                      Mean  : 54.943
##                      3rd Qu.: 72.250
##                      Max.   :374.000
```

365个城市的PM2.5指数的平均值是54.94，中位数是47.13，最小值是2.67，最大值是374

- 城市PM2.5平均值的数值分布

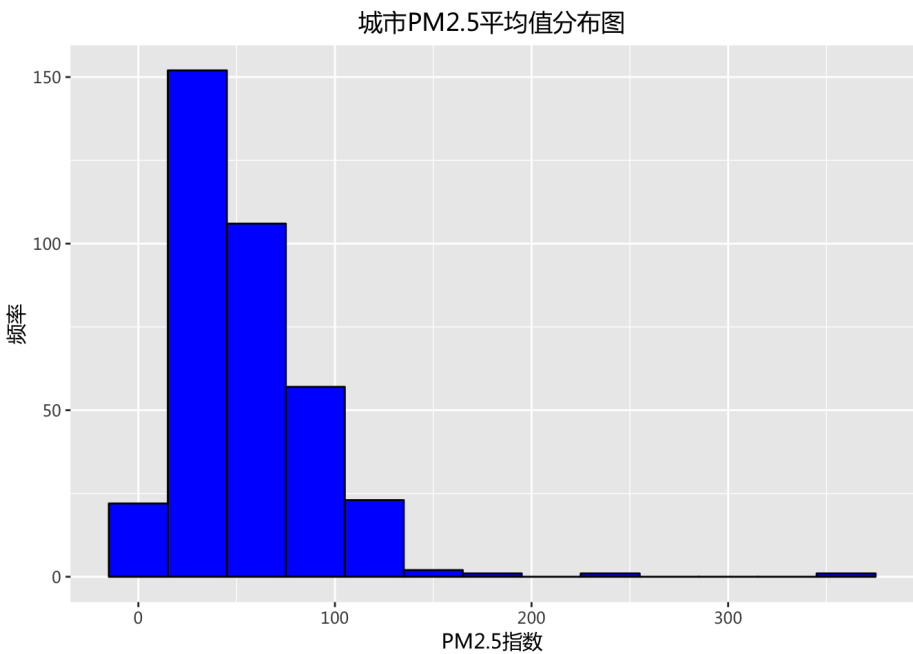
```
avg_city_pm2.5 %>%
  count(PM2.5指数 = cut_width(城市PM2.5平均值, 30))
```

PM2.5指数	n
<fctr>	<int>
[-15,15]	22
(15,45]	152
(45,75]	106
(75,105]	57
(105,135]	23
(135,165]	2
(165,195]	1
(225,255]	1
(345,375]	1

9 rows

- 绘制城市PM2.5平均值的直方图

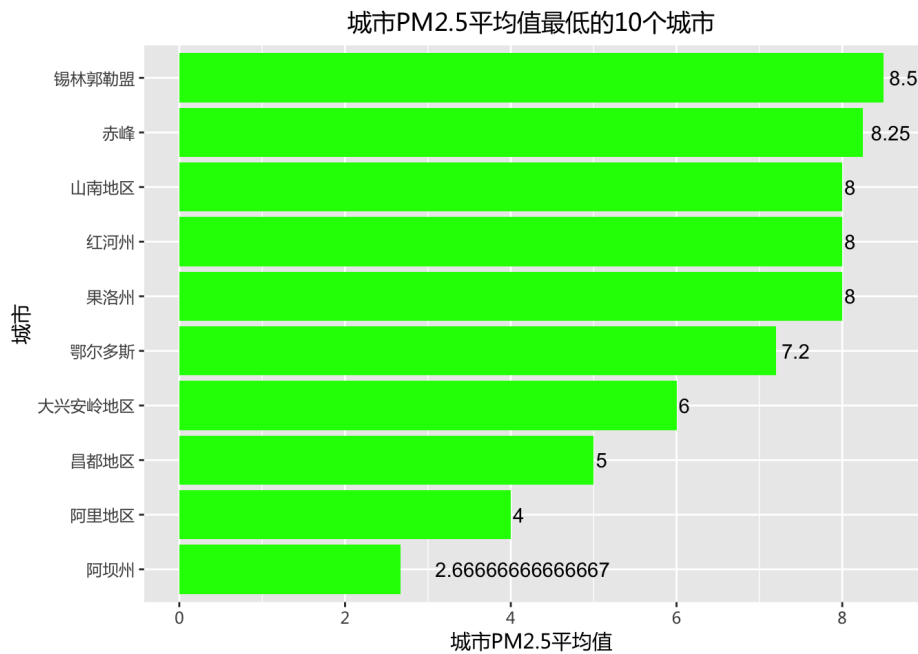
```
ggplot(avg_city_pm2.5, aes(城市PM2.5平均值)) +
  geom_histogram(color="black", fill="blue", binwidth = 30) +
  labs(title="城市PM2.5平均值分布图", x="PM2.5指数", y="频率") +
  plot_theme
```



有360个城市的PM2.5指数在0-135之间，占总体的98.6%，分布与城市AQI指数基本相同，显示两者呈正相关关系

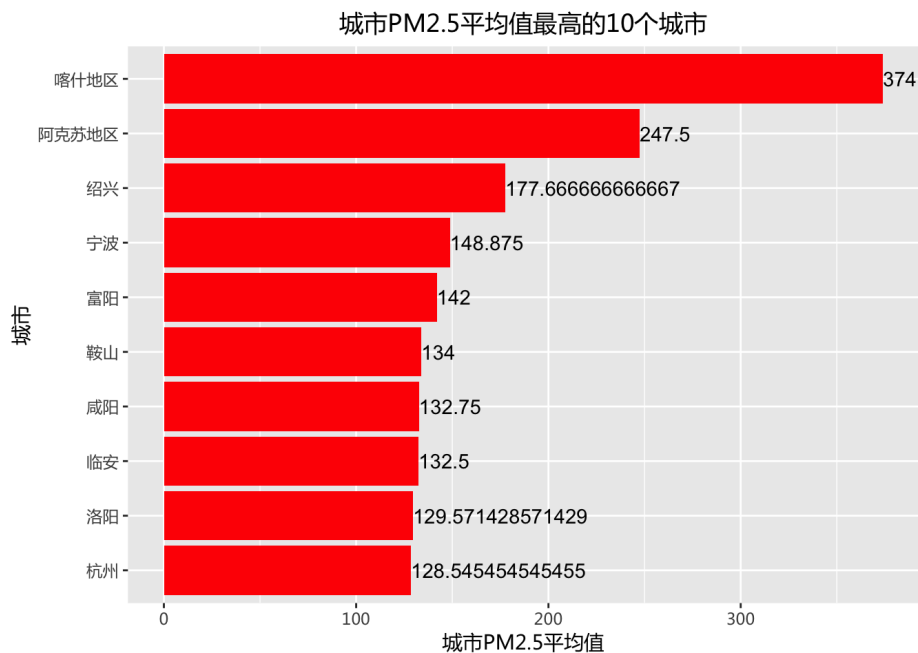
- 城市PM2.5平均值最低的10个城市

```
avg_city_pm2.5 %>%
  arrange(城市PM2.5平均值) %>%
  head(10) %>%
  ggplot(aes(reorder(城市, 城市PM2.5平均值), 城市PM2.5平均值)) +
  geom_bar(stat = "identity", fill="green") +
  labs(title = "城市PM2.5平均值最低的10个城市", x="城市", y="城市PM2.5平均值") +
  geom_text(aes(label=城市PM2.5平均值), hjust=-0.2) +
  coord_flip() +
  plot_theme
```



- 城市PM2.5平均值最高的10个城市

```
avg_city_pm2.5 %>%
  arrange(desc(城市PM2.5平均值)) %>%
  head(10) %>%
  ggplot(aes(reorder(城市, 城市PM2.5平均值), 城市PM2.5平均值)) +
  geom_bar(stat = "identity", fill="red") +
  labs(title = "城市PM2.5平均值最高的10个城市", x="城市", y="城市PM2.5平均值") +
  geom_text(aes(label=城市PM2.5平均值), hjust=0) +
  coord_flip() +
  plot_theme
```



5. 分析PM10变量

- 查看PM10的描述统计量

```
describe(aqi$PM10)
```



```
## aqi$PM10
##      n missing distinct      Info      Mean      Gmd      .05      .10
## 1389      64      229         1    97.95    61.07    30.0    38.0
##   .25   .50   .75   .90   .95
## 57.0   87.0  125.0  169.2  197.0
##
## lowest :    1    5    7    8    9, highest: 386  714  801  955 1135
```

PM10变量有1389个值，有64个缺失值，其中平均值是97.95，中位数是87，最小值是1，最大值是1135

- 按城市分组，计算各城市的PM10平均值

```
avg_city_pm10 <- aqi %>%
  group_by(城市) %>%
  summarise(城市PM10平均值 = mean(PM10))
avg_city_pm10
```

城市	城市PM10平均值
<chr>	<dbl>
阿坝州	NA
阿克苏地区	1045.00000
阿拉善盟	37.33333
阿勒泰地区	53.00000
阿里地区	8.50000
安康	99.33333
安庆	143.00000
安顺	40.75000
安阳	96.40000
鞍山	235.00000

1-10 of 365 rows

Previous 1 2 3 4 5 6 ... 37 Next

- 查看城市PM10平均值的描述统计量

```
summary(avg_city_pm10)
```

```
##      城市      城市PM10平均值
## Length:365      Min.   :  8.50
## Class :character 1st Qu.: 59.00
## Mode  :character Median : 85.00
##                      Mean  : 98.81
##                      3rd Qu.:123.45
##                      Max.  :1045.00
##                      NA's   :46
```

365个城市里，有46个城市没有数据，余下的319个城市里，PM10指数的平均值是98.81，中位数是85，最小值是8.5，最大值是1045

- 城市PM10平均值的数值分布

```
avg_city_pm10 %>%
  count(PM10指数 = cut_width(城市PM10平均值, 30))
```

```
## Warning: Factor `PM10指数` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

PM10指数	n
<fctr>	<int>
[-15,15]	1
(15,45]	38
(45,75]	96
(75,105]	74
(105,135]	48
(135,165]	28

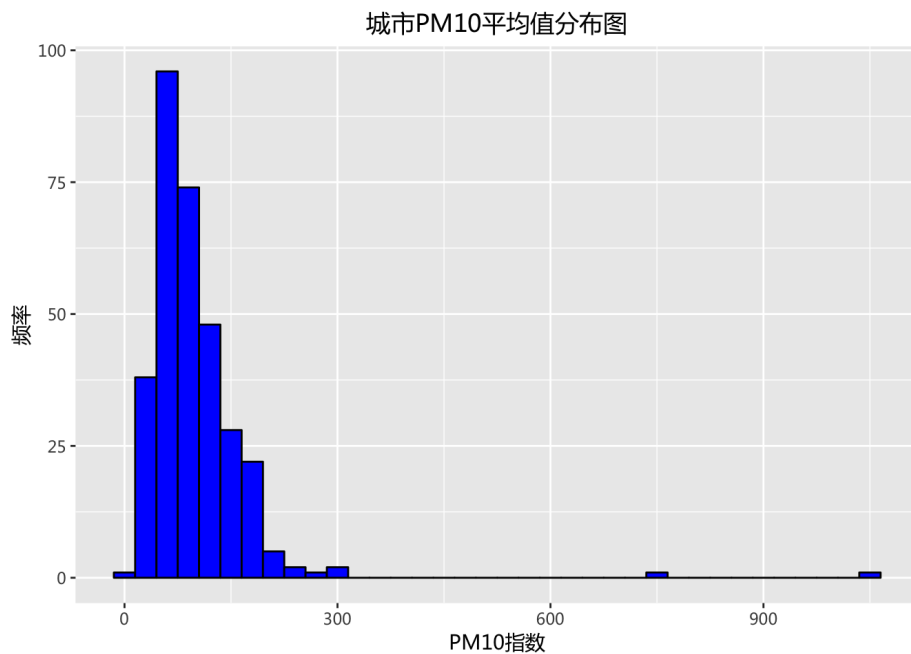
PM10指数 <fctr>	n <int>
(165,195]	22
(195,225]	5
(225,255]	2
(255,285]	1

1-10 of 14 rows

Previous 1 2 Next

- 绘制城市PM10平均值的直方图

```
ggplot(avg_city_pm10, aes(城市PM10平均值)) +
  geom_histogram(color="black", fill="blue", binwidth = 30, na.rm = TRUE) +
  labs(title = "城市PM10平均值分布图", x="PM10指数", y="频率") +
  plot_theme
```

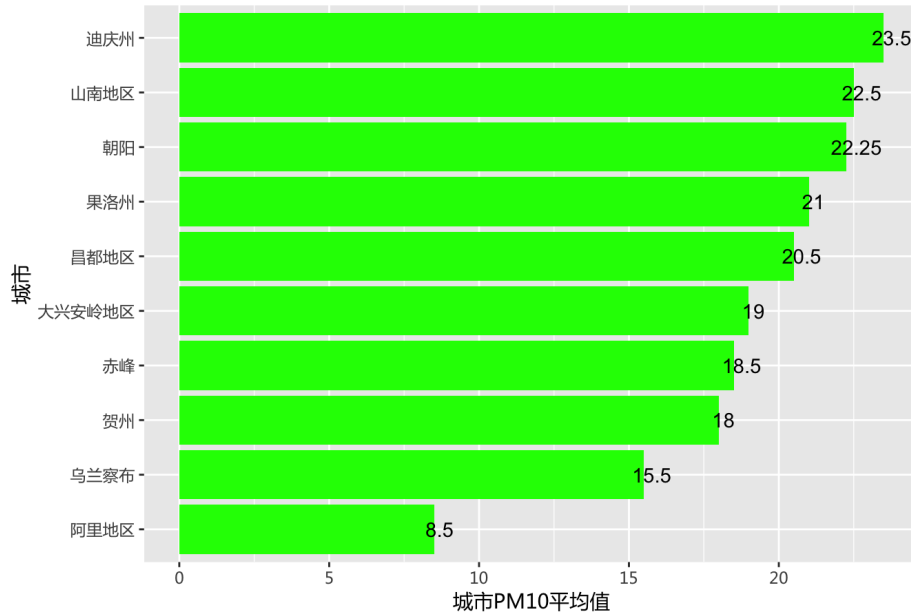


有307个城市的PM10指数在0-195之间，占总体的96.2%，分布与城市AQI指数和PM2.5指数基本相同，显示三者呈现正相关关系

- 城市PM10平均值最低的10个城市

```
avg_city_pm10 %>%
  arrange(城市PM10平均值) %>%
  head(10) %>%
  ggplot(aes(reorder(城市, 城市PM10平均值), 城市PM10平均值)) +
  geom_bar(stat = "identity", fill="green") +
  labs(title = "城市PM10平均值最低的10个城市", x="城市", y="城市PM10平均值") +
  geom_text(aes(label=城市PM10平均值), hjust=0.3) +
  coord_flip() +
  plot_theme
```

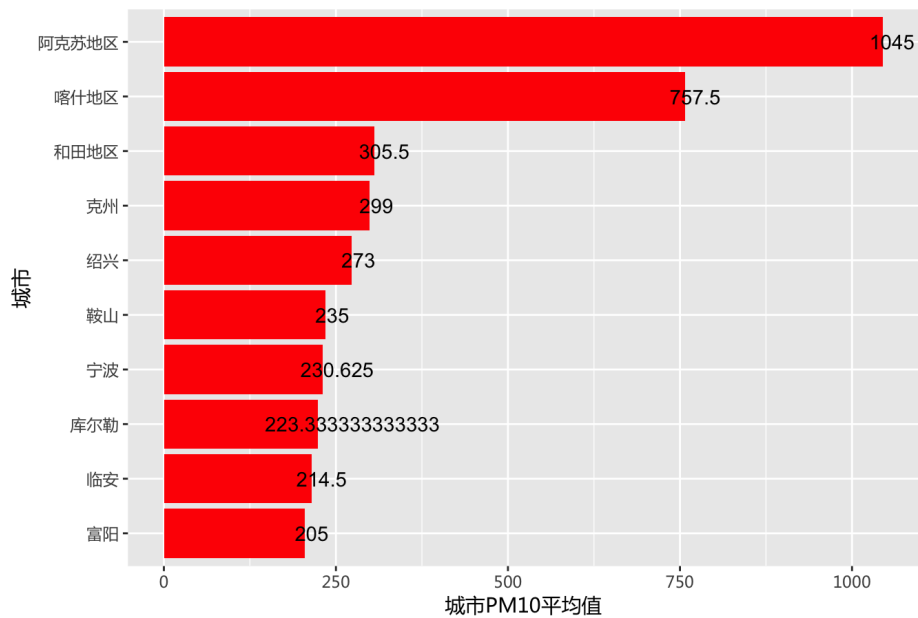
城市PM10平均值最低的10个城市



- 城市PM10平均值最高的10个城市

```
avg_city_pm10 %>%
  arrange(desc(城市PM10平均值)) %>%
  head(10) %>%
  ggplot(aes(reorder(城市, 城市PM10平均值), 城市PM10平均值)) +
  geom_bar(stat = "identity", fill="red") +
  labs(title = "城市PM10平均值最高的10个城市", x="城市", y="城市PM10平均值") +
  geom_text(aes(label=城市PM10平均值), hjust=0.3) +
  coord_flip() +
  plot_theme
```

城市PM10平均值最高的10个城市



6. 分析首要污染物变量

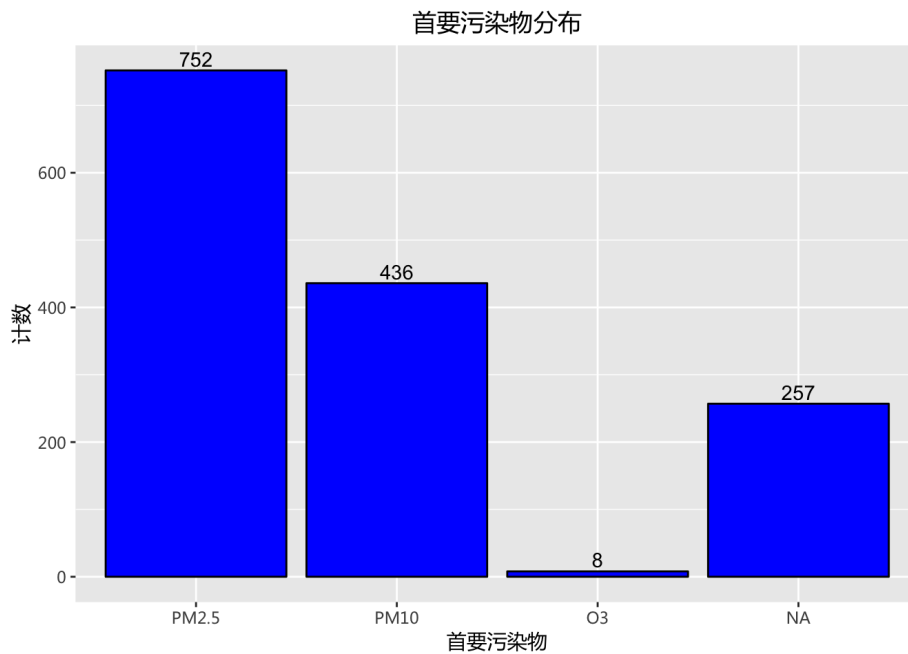
- 查看首要污染物的描述统计量

```
summary(aqi$首要污染物)
```

```
## PM2.5 <NA> PM10 O3
## 752 257 436 8
```

- 绘制首要污染物的条形图

```
ggplot(aqi, aes(aqi$首要污染物)) +
  geom_bar(color="black", fill="blue") +
  labs(title="首要污染物分布", x="首要污染物", y="计数") +
  geom_text(aes(label=as.character(..count..)), stat="count", vjust=-0.3) +
  plot_theme
```



在首要污染物中，有752个地区是PM2.5，有436个地区是PM10，有8个地区是臭氧（O3），有257个缺失值（NA），这些缺失值所代表的都是污染等级为优的地区

7. 分析污染等级变量

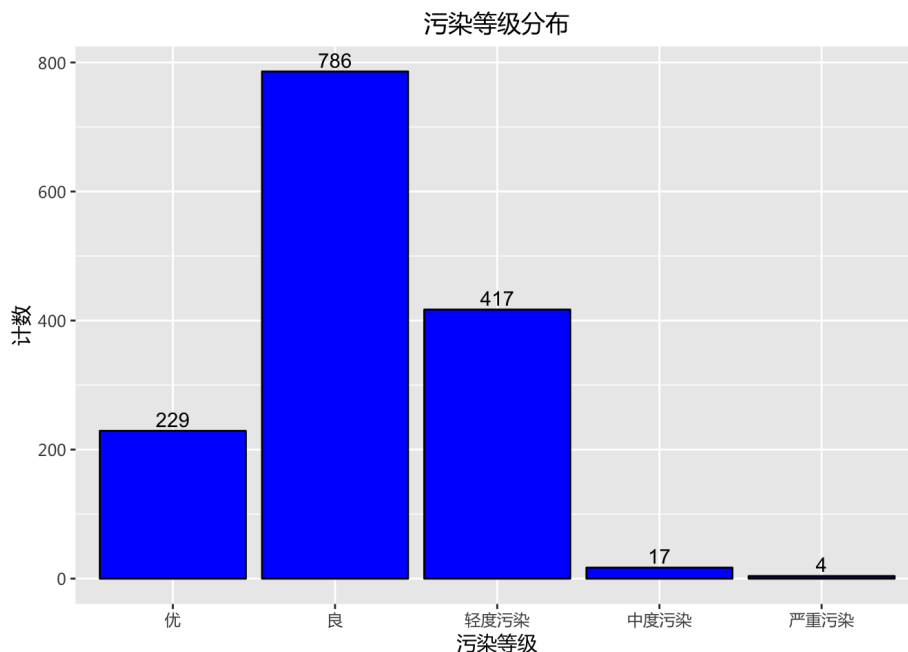
- 查看污染等级的描述统计量

```
summary(aqi$污染等级)
```

```
##      优      良 轻度污染 中度污染 严重污染
##      229     786     417      17        4
```

- 绘制污染等级的条形图

```
ggplot(aqi, aes(aqi$污染等级)) +
  geom_bar(color="black", fill="blue") +
  labs(title = "污染等级分布", x="污染等级", y="计数") +
  geom_text(aes(label = as.character(..count..)), stat = "count", vjust=-0.3) +
  plot_theme
```



在污染等级中，有229个地区为优，有786个地区为良，有417个地区为轻度污染，有17个地区为中度污染，有4个地区为严重污染

5.2 多因子探索性数据分析

参考资料

1. 城市空气质量等级
(<https://baike.baidu.com/item/%E5%9F%8E%E5%B8%82%E7%A9%BA%E6%B0%94%E8%B4%A8%E9%87%8F%E7%AD%89%E7%BA%A7/8fr=aladdin>)