aqi

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(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")

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") # 转换为因子
```

- 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

```
describeBy(aqi)
```

	vars <int>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>
城市*	1	1453	NaN	*NA*	*NA*	NaN	*NA*	Inf	-Inf
地区*	2	1453	NaN	*NA*	*NA*	NaN	*NA*	Inf	-Inf
城市AQI	3	1453	85.528562	42.6648647	76	80.960447	34.0998	26	500
PM2.5	4	1453	57.835513	36.8658286	50	54.138435	31.1346	1	476
PM10	5	1389	97.948884	68.1874396	87	90.754717	48.9258	1	1135
首要污染物*	6	1196	1.793531	0.8930629	1	1.735168	0.0000	1	4
污染等级*	7	1453	2.161046	0.7004107	2	2.179708	0.0000	1	5
地区AQI	8	1453	85.904336	45.2100828	75	80.766982	34.0998	11	500

8 rows | 1-10 of 14 columns

> *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

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

```
##      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
```

> *数据集总共有365个城市。*

2. 分析地区变量

- 统计地区数量

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

```
##      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
```

> *收集空气数据的地区有1264个。整个数据集有1453个观测，可以看出其中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

- 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
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，最小值是26，最大值是500。*

- 城市AQI平均值的数值分布

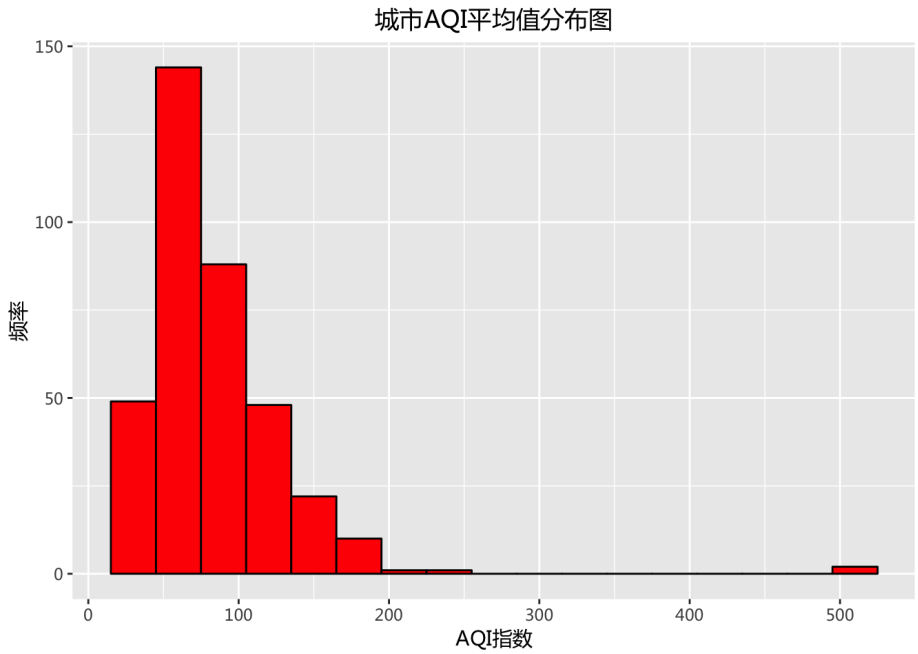
```
avg_city_aqi %>%
  count(AQI指数 = cut_width(城市AQI平均值, 30))
```

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

```
ggplot(avg_city_aqi, aes(城市AQI平均值)) +
  geom_histogram(color="black", fill="red", binwidth = 30) +
  labs(title="城市AQI平均值分布图", x="AQI指数", y="频率") +
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(family = "MicrosoftYaHei"))
```



有个49城市的AQI指数在15-45之间，等级为优，占比为13.4%；

有232个城市的AQI指数在45-105之间，等级为良，占比为63.6%；

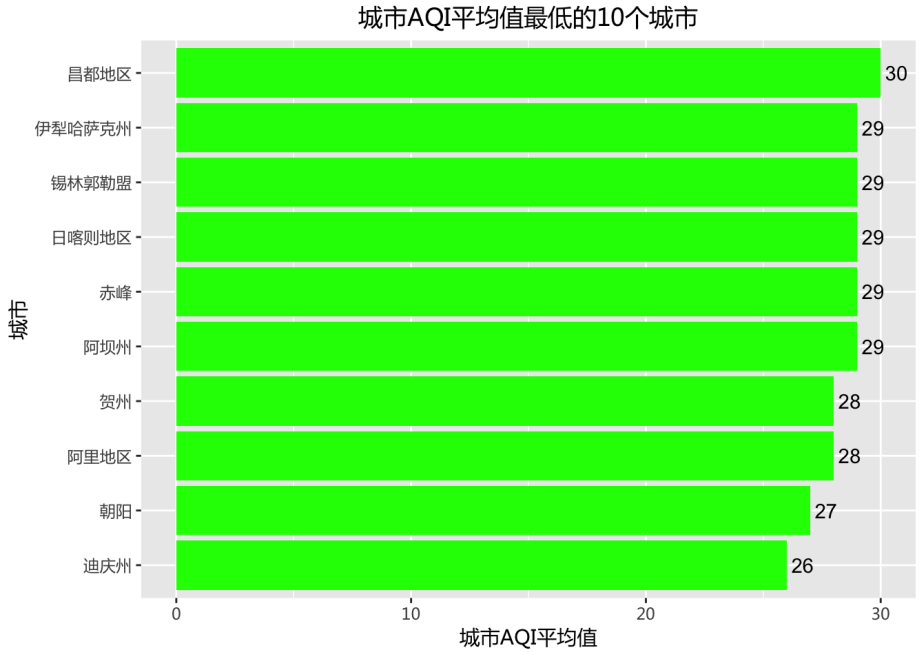
有80个城市的AQI指数在105-195之间，等级为轻度污染，占比为21.9%；

有2个城市的AQI指数在195-255之间，等级为中度污染，占比为0.5%；

有2个城市的AQI指数在495-525之间，等级为重度污染，占比为0.5%。

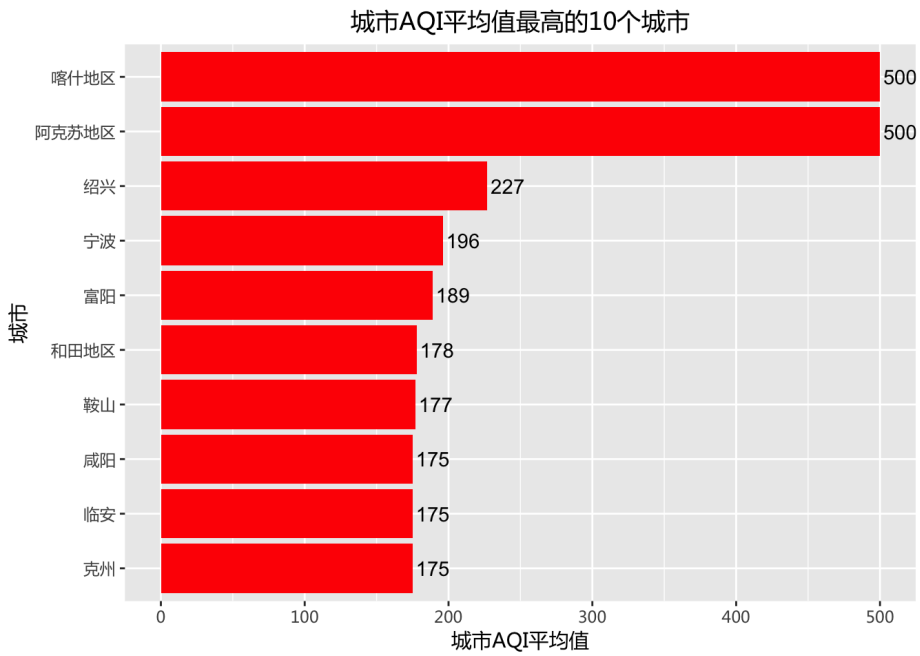
- 城市AQI平均值最低的10个城市

```
avg_city_aqi %>%
  arrange(城市AQI平均值) %>%  # 按从小到大升序排列
  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") +
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(family = "MicrosoftYaHei")) +
  geom_text(aes(label=城市AQI平均值), hjust=-0.2) +
  coord_flip()
```



- 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") +
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(family = "MicrosoftYaHei")) +
  geom_text(aes(label=城市AQI平均值), hjust=-0.1) +
  coord_flip()
```



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%A7/8
fr=aladdin)