--- title: "链家数据分析" CJKmainfont: Songti SC author: "姚文喆" date: "2023.10.20" output: html_document: code_folding: show fig_caption: yes fig_width: 10 highlight: tango number_sections: yes theme: cosmo toc: yes word_document: toc: yes pdf_document: latex_engine: xelatex --- ```{r setup, include = FALSE,echo = FALSE} knitr::opts_chunk$set(echo = FALSE,error = FALSE, warning = FALSE, message = FALSE, out.width = "100%", split = FALSE, fig.align = "center") #load library library(tidyverse) library(kableExtra) library(lubridate) library(scales) library(plotly) library(patchwork) library(ggrepel) library(showtext) showtext_auto(enable=TRUE) ``` # 你的主要发现 1. 房屋近地铁价格比非近地铁每平方米价格要高 1. 房屋楼层越高每平方米价格越高 1. 房屋房本满两年价格更优惠 # 数据介绍 本报告**链家**数据获取方式如下： 报告人在2023年9月12日获取了[链家武汉二手房网站](https://wh.lianjia.com/ershoufang/)数据。 - 链家二手房网站默认显示100页，每页30套房产，因此本数据包括3000套房产信息； - 数据包括了页面可见部分的文本信息，具体字段及说明见作业说明。 **说明：**数据仅用于教学；由于不清楚链家数据的展示规则，因此数据可能并不是武汉二手房市场的随机抽样，结论很可能有很大的偏差，甚至可能是错误的。 ```{r} # 载入数据和预处理 lj<- read_csv("/Users/macbook/Desktop/2023-09-12_cleaned.csv") # EDA ---------- ---------- ## 如下语句可以解决画图中的中文显示问题，当然你可以用showtext包来解决 theme_set(theme(text = element_text(family="Songti SC",size = 10))) #这里family设置成你系统中的中文字体名。 # 做一些数据预处理，比如把字符型变成factor。 ``` # 数据概览 数据表（lj)共包括`r names(lj)`等`r ncol(lj)`个变量,共`r nrow(lj)`行。表的前10行示例如下： ```{r} lj %>% head(10) %>% kable(caption = "武汉链家二手房") %>% kable_styling() ``` 各变量的简短信息： ```{r} glimpse(lj) ``` 各变量的简短统计： ```{r} summary(lj) ``` 可以看到： - 此数据中房价的均值为15，148.00元/平方 - 此数据中房型为三室两厅的居多，平均面积为100平方米。 # 探索性分析 ## 变量1的数值描述与图形 发现： - 房屋近地铁比非近地铁每平方米价格要高 ```{r} lj_ns<- filter(lj,near_subway%in%c("近地铁",NA)) lj_ns<-group_by(lj_ns,near_subway) lj_mns<- summarise(lj_ns,mean_near_subway=mean(price_sqm)) view(lj_mns) ggplot(data=lj_ns,mapping = aes(x=price_sqm))+ geom_histogram(binwidth = 2000)+facet_wrap(~near_subway) ``` ## 变量2的数值描述与图形 发现： - 房屋楼层越高每平方米价格越高 ```{r} lj_l<-filter(lj,property_t_height%in%c(0,10)) lj_m<- filter(lj,property_t_height%in%c(11,20)) lj_h<-filter(lj,property_t_height>20) lj_ml<- summarise(lj_l,mean_low=mean(price_sqm)) lj_mm<- summarise(lj_m,mean_mid=mean(price_sqm)) lj_mh<- summarise(lj_h,mean_high=mean(price_sqm)) ggplot(data=lj_l,mapping = aes(x=price_sqm))+ geom_histogram(binwidth = 1000)+labs(title = "低区房价柱状图") ggplot(data=lj_m,mapping = aes(x=price_sqm))+ geom_histogram(binwidth = 1000)+labs(title = "中区房价柱状图") ggplot(data=lj_h,mapping = aes(x=price_sqm))+ geom_histogram(binwidth = 1000)+labs(title = "高区房价柱状图") ``` ## 变量...的数值描述与图形 发现： - 房屋房本满两年价格更优惠 ```{r} lj_if<-filter(lj,if_2y%in%c("房本满两年",NA)) lj_if<-group_by(lj_ns,if_2y) lj_mif<- summarise(lj_if,mean_if_2y=mean(price_sqm)) view(lj_mif) ggplot(data=lj_if,mapping = aes(x=price_sqm))+ geom_histogram(binwidth = 2000)+facet_wrap(~if_2y) ``` ## 探索问题1 发现： - 房屋单价是否会随着房屋面积的增加而增加 ```{r} lj_1<-filter(lj,building_area<=400) ggplot(data=lj_1,mapping = aes(x=building_area,y=price_sqm))+ geom_point()+geom_smooth(se=FALSE) ``` ## 探索问题2 发现： - 房屋单价是否会随着楼层增加而增加 ```{r} ggplot(data=lj_1,mapping = aes(x=property_t_height,y=price_sqm))+ geom_point()+geom_smooth(se=FALSE) ``` # 发现总结 影响房价的因素有很多种，区位、面积、楼层和地理位置等等，只有进行综合分配考虑才能寻找到性价比最高的房屋。