به‌نام خدا

گزارش تمرین عملی سری ۵

محمدحسین کلهرجهاندوست

۹۵۱۰۹۹۶۱

برای بدست آوردن منطقه با بیشترین آلودگی، ابتدا کل داده را، با دستور group_by بر حسب منطقه گروه بندی می‌کنیم. سپس با استفاده از دستور summarise و پارامتر mean، میانگین آلودگی هر گروه را بدست می‌آوریم. سپس با استفاده از دستور arrange و پارامتر desc، مناطق را بر حسب میانگین آلودگی‌شان، به صورت نزولی مرتب می‌کنیم. بدین ترتیب سطر اول جدول، همان آلوده‌ترین منطقه بر اساس معیار میانگین است. بنابراین با استفاده از دستور slice(1:1) اولین سطر جدول را برمی‌گردانیم.
برای کشیدن نمودار ۲۰ منطقه ای که بیشترین آلودگی را دارند، طبق قسمت قبل عمل می‌کنیم. این بار با استفاده از دستور slice(1:20)، ۲۰ سطر اول را می‌گیریم و به عنوان ورودی دیتای تابع ggplot می‌دهیم. محور x را نام مناطق و محور y را میانگین آلودگی مناطق، مقداردهی می‌کنیم. سپس با استفاده از دستور geom_bar نمودار ستونی را رسم می‌کنیم. برای ایالت‌ها نیز همین کار را تکرار می‌کنیم.

ابتدا متغیر Cateegory را با دستور as.Factor() به متغیر categorical تبدیل می‌کنیم. البته بر اساس مشاهده‌ی شخصی، اگر این دستور را هم اجرا نمی‌کردیم، خود R، آن را categorical در نظر می‌گرفت. سپس دیتافریم aqi را به عنوان پارامتر دیتای تابع ggplot ورودی می‌دهیم. همچنین محور ستون Category دیتافریم aqi را به عنوان محور x و میزان آلودگی (ستون AQI دیتافریم aqi) را به عنوان محور y قرار می‌دهیم. این کار را به کمک (aqi$Category) و (aqi$AQI انجام می‌دهیم. سپس نمودار جعبه ای را با کمک تابع geom_boxplot() رسم می‌کنیم.

ابتدا میانگین آلودگی در هر روز را بدست می‌آوریم. برای این کار کل داده را با دستور group_by بر حسب روز گروه بندی می‌کنیم. سپس با استفاده از دستور summarise و پارامتر mean، میانگین آلودگی هر گروه (روز) را بدست می‌آوریم. این دیتا را به عنوان ورودی data تابع ggplot می‌دهیم. روز ها را به عنوان محور x و میانگین آلودگی در هر روز (که در خط قبلی بدست آورده بودیم) را به عنوان محور y قرار می‌دهیم. سپس با استفاده از تابع geom_line نمودار خطی مورد نظر را رسم می‌کنیم.

باید ببینیم هر کدام از آلاینده ها چند بار به عنوان مهم ترین عامل آلودگی هوای واشنگتن، تکرار شده اند. برای این کار ابتدا با استفاده از دستور select دو ستون نام منطقه و مهم ترین عامل آلودگی را انتخاب می‌کنیم. سپس با استفاده از دستور filter داده‌های مربوز به منطقه‌ی washington رو جدا می‌کنیم. حالا با استفاده از دستور group_by داده‌ها را بر حسب نوع آلاینده گروه‌بندی می‌کنیم. در نهایت با استفاده از دستور count() تعداد تکرار هر آلاینده را حساب می‌کنیم. با استفاده از دستور arrange تعداد تکرارها را به صورت نزولی مرتب می‌کنیم و سطر اول را به عنوان خروجی نشان می‌دهیم. همانطور که از خروجی و نمودار قسمت بعد مشخص است، آلاینده‌ی Ozone با ۲۲۷۸ بار تکرار مهم ترین آلاینده‌ی هوای واشنگتن است.

برای کشیدن نمودار همان روند بالا را تکرار می‌کنیم. در نهایت با استفاده از دستور ggplot با قرار دادن آلاینده‌ها به عنوان محور x و تعداد تکرار (n) به عنوان محور y و دستور geom_bar نمودار ستونی را رسم می‌کنیم. دقت کنید ورودی پارامتر دیتای تابع ggplot همان جدول نام آلاینده‌ها با تعداد تکرار یعنی خروجی تابع count() است.

برای بدست آوردن یک بازه‌ی اطمینان برای میانگین آلودگی هر منطقه باید میانگین آلودگی هر منطقه، انحراف معیار هر منطقه و تعداد تکرار هر منطقه را داشته باشیم. برای این کار ابتدا با استفاده از دستور group_by، داده را بر حسب نام مناطق گروه‌بندی می‌کنیم. بنابراین grouped یک دیتافریم گروه‌بندی شده بر حسب مناطق است. حالا با استفاده از دستور summarise، میانگین، انحراف معیار و تعداد تکرار را بدست می‌آوریم و به ترتیب در بردار m، sd و n می‌ریزیم. حالا با توجه به فرمول بازه اطمینان ابتدا و انتهای هر بازه را بدست می‌آوریم و به ترتیب در بردار start و end می‌ریزیم. حالا با استفاده از دستور data.frame() دیتافریمی می‌سازیم که ستون اول آن نام مناطق، ستون دوم آن بردار start و ستون سوم آن بردار end است.

با دو دیدگاه این سوال را بررسی می‌کنیم. ابتدا به کمک دستور filter داده‌های مربوط به زمستان (winter) و تابستان (summer) را بدست می‌آوریم. برای محاسبه‌ی p-value، ستون AQI داده‌های تابستان و زمستان را به تابع t.test() ورودی می‌دهیم. فرض صفر را این می‌گیریم که میانگین آلودگی در تابستان با میانگین آلودگی در زمستان برابر است. با توجه به میانگین آلودگی در تابستان و زمستان، فرض مقابل برابر با این است که میانگین آلودگی در تابستان بیشتر از

به‌نام خدا

گزارش تمرین عملی سری ۵

محمدحسین کلهرجهاندوست

۹۵۱۰۹۹۶۱

میانگین آلودگی در زمستان است. پس مقدار پارامتر alternative را greater سِت می‌کنیم. همچنین conf.level را ۹۵٪ در نظر می‌گیریم. با توجه به این که p-value بسیار نزدیک به صفر است، فرض صفر را قویا رد می‌کنیم و فرض مقابل را می‌پذیریم. یعنی آلودگی در تابستان بیشتر یا مساوی آلودگی در زمستان است.

رویکردی دیگر، استفاده از نمونه‌ی جفت‌شده است. یعنی می‌توانیم میانگین آلودگی هر منطقه در تابستان و زمستان را محاسبه کنیم. سپس با فرض صفر سوال قبل، p-value را محاسبه کنیم. البته بعد از انجام این کار، مشاهده شد که برای ۱۱ منطقه در زمستان اطلاعاتی ثبت نشده است. بنابراین خود R بعد از اعمال دستور merge، این ۱۱ شهر را از نمونه‌ها حذف می‌کند. باید توجه کنیم برای استفاده از نمونه‌ی جفت‌شده باید پارامتر paired را TRUE کنیم تا جفت‌شده محاسب کند. با این روش نیز p-value بسیار کوچک می‌شود و فرض صفر رد می‌شود و می‌توان نتیجه‌گیری کرد که آلودگی در تابستان بیشتر یا مساوی زمستان است.

در این سوال نیز مانند سوال قبل عمل می‌کنیم و همان فرض صفر را می‌گیریم. تفاوتی که این سوال با سوال قبل دارد این است که، چون در صورت سوال می‌خواهد ادعای «تفاوت داشتن آلودگی در زمستان و تابستان» را بررسی کنیم، بنابراین فرض مقابل این می‌شود که این دو میانگین برابر نیست. پس پارامتر alternative را two.sided سِت می‌کنیم. چون p-value خیلی کوچک می‌شود، فرض صفر را رد می‌کنیم. یعنی میانگین تابستان و زمستان با هم تفاوت دارند.

ابتدا با استفاده از دستور filter داده‌های مربوط به ایالت‌های تگزاس و میشیگان را جدا می‌کنیم به ترتیت در txs و mch می‌ریزیم. سپس ستون‌های مربوط به AQI را به ترتیب در txsAQI و mchAQI می‌ریزیم. حالا فرض صفر را این در نظر می‌گیریم که میانگین آلودگی تگزاس و میشیگان با هم برابر است. فرض مقابل هم به صورت پیش فرض two.sided در نظر می‌گیرم. با توجه به مقدار کوچک p-value، فرض صفر را رد می‌کنیم. یعنی میانگین آلودگی تگزاس و میشیگان تفاوت معنادار دارند. این نتیجه با توجه به اختلاف کم میانگین این دو ایالت، جالب است. نشان می‌دهد که این نزدیکی اتفاقی است.
اگر فرض مقابل را به less تغییر دهیم مجددا به p-value بسیار کوچک می‌رسیم و فرض صفر را رد می‌کنیم و فرض مقابل یعنی میانیگین آلودگی تگزاس کمتر از میشیگان است را می‌پذیریم.

ابتدا با استفاده از دستور filter سه ایالت آلاباما، کالیفرنیا و نیویورک را جدا می‌کنیم. سپس با استفاده از دستور aov آماره‌ی F و p-value را حساب می‌کنیم. فرض صفر این است که میانگین آلودگی این سه شهر تفاوت ندارد. با توجه به p-value بسیار کوچک فرض صفر رد می‌شود و این میانگین آلودگی این سه شهر تفاوت معنادار دارند. آماره‌ی F برابر با ۲۵۵۷ است.

در سوال ۱، هاوایی به عنوان آلوده‌ترین منطقه با معیار میانگین آلودگی مشخص شد. حالا نمودار خطی آلودگی بر حسب روز را رسم کردیم.
در سوال ۱، کالیفرنیا به عنوان آلوده‌ترین ایالت مشخص شد. حالا نمودار ستونی میانگین آلودگی مناطق مختلف کالیفرنیا را رسم کردیم.
در سوال ۱، هاوایی به عنوان آلوده‌ترین منطقه با معیار میانیگین آلودگی مشخص شد. حالا نمودار ستونی تعداد تکرار نوع هوا بر حسب نوع هوا را رسم می‌کنیم.