Data-Driven decision-making class - Spring 1401

Exercise #1

Healthcare issue

Mahsa Choopannezhad – student number: 98207477

Overview

تمرین اول درس مدلسازی و تصمیم گیری داده محور، بیشتر تمرکز تمرین روی مصورسازی و آزمون فرض بر روی یک دیتاست است.

Goals

دیتاست شامل106,987 رکورد و 14 ستون است که اطلاعاتی راجع به مراجعین به یک مرکز درمانی را شامل می‌شود. هدف این است که با بررسی دتیاست بتوانیم به اقدامات و راهکارهایی برسیم که نرخ عدم مراجعه را کاهش دهد و موجب کاهش هدررفت منابع این مرکز درمانی شود.

Questions and results

Introduction to the dataset:

ابتدا سعی می‌شود که هر یک از ستون های دیتاست بررسی شود و متغیرهای کتگوریکال را فکتور میکنیم.

این قسمت از کد که با هدر #----------Data Cleaning---------- مشخص شده است چند چک پوینت را برای تمیز کردن دیتاست بررسی میکند.

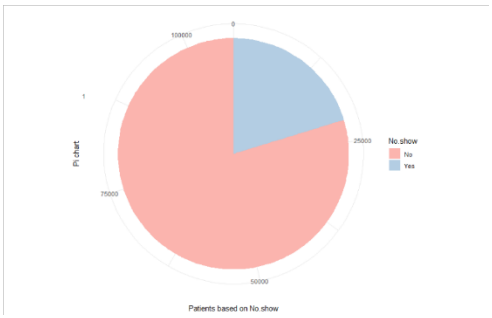
در ادامه به نظر میرسه که باتوجه به خلاصه دیتاست، سن منفی منطقی نیست وبا بررسی تعداد رکوردهای با سن منفی (که ۱ رکورد است) میتوان آن را خذف کرد. چک پوینت دیگه ای که باتوجه به این دتیاست میشه در نظر گرفت، تکراری بودن appointment id هست که هیچ رکورد تکراری از منظر appointment id نداریم.

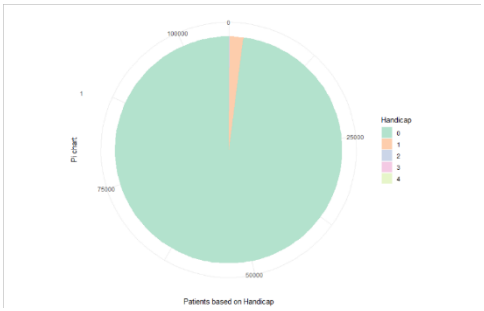
یکی دیگه از مواردی که هست این هستش که متغیر handicap که در شرح تمرین به عنوان متغیر باینری معرفی شده اینگونه نیست و دارای ۵سطح هست که به نظر میرسه میزان معلولیت را مشخص میکند.

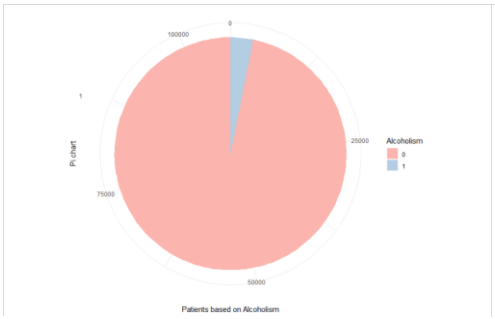
Question:1#

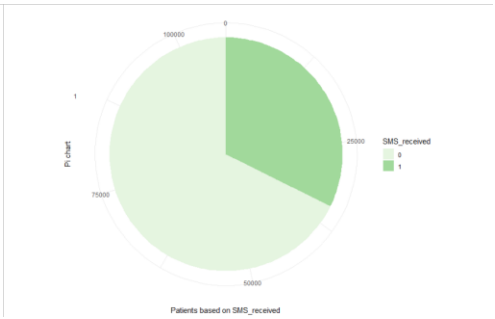
با رسم نمودارهای مناسب برخی از ویژگی های متغیرها را شناسایی کنید. برای این شناسایی حداقل ۵ نمودار با استفاده از ggplot۲ رسم کنید و برای هر نمودار توضیح دهید که چه چیزی در مورد پدیده show-no متوجه می شوید. سعی کنید نمودارهای شما کامل و حرفه‌ای باشد.

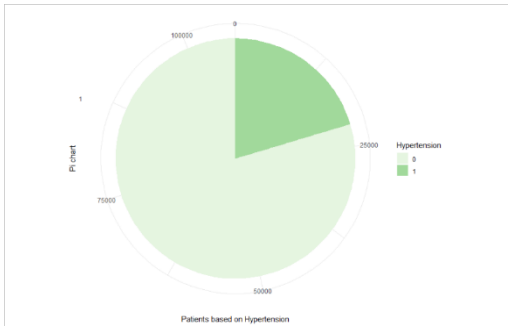
قبلا از اینکه وارد بررسی ارتباط متغیر no.show و بقیه متغیرها بشیم، توزیع جمعیتی نمونه‌ها را روی متغیرهای مختلف با pi chart های مختلف بررسی میکنیم. نکته قابل توجه در این توزیع ها این است که بیماری فشار خون بیشتر از بقیه بین جمعیت مراجعه کننده شایع است.

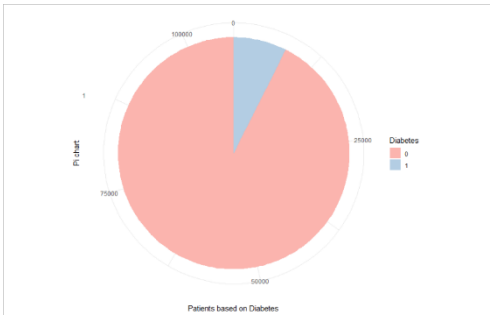


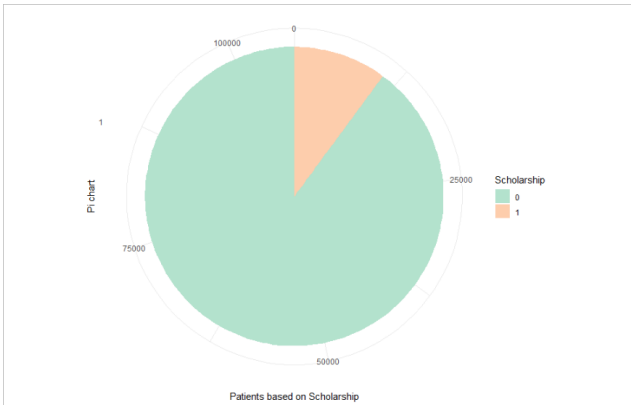








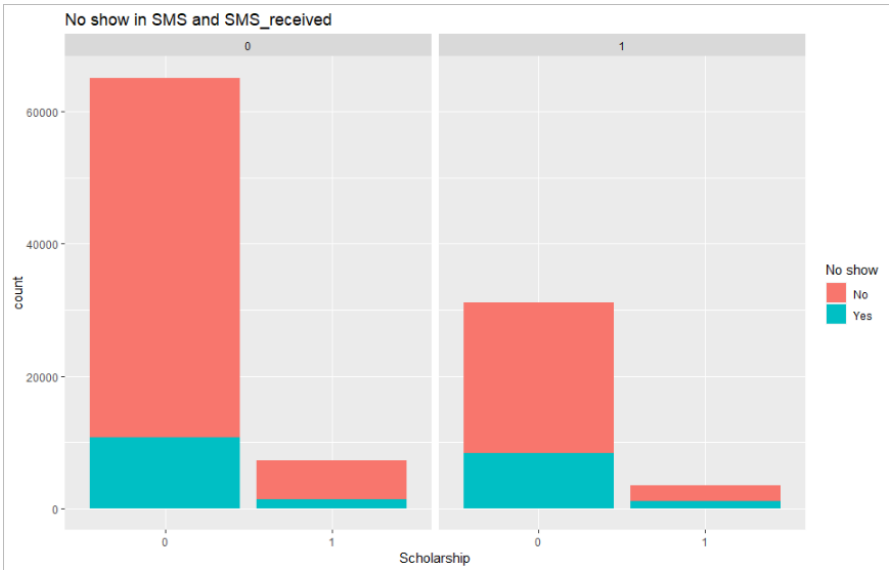




A:

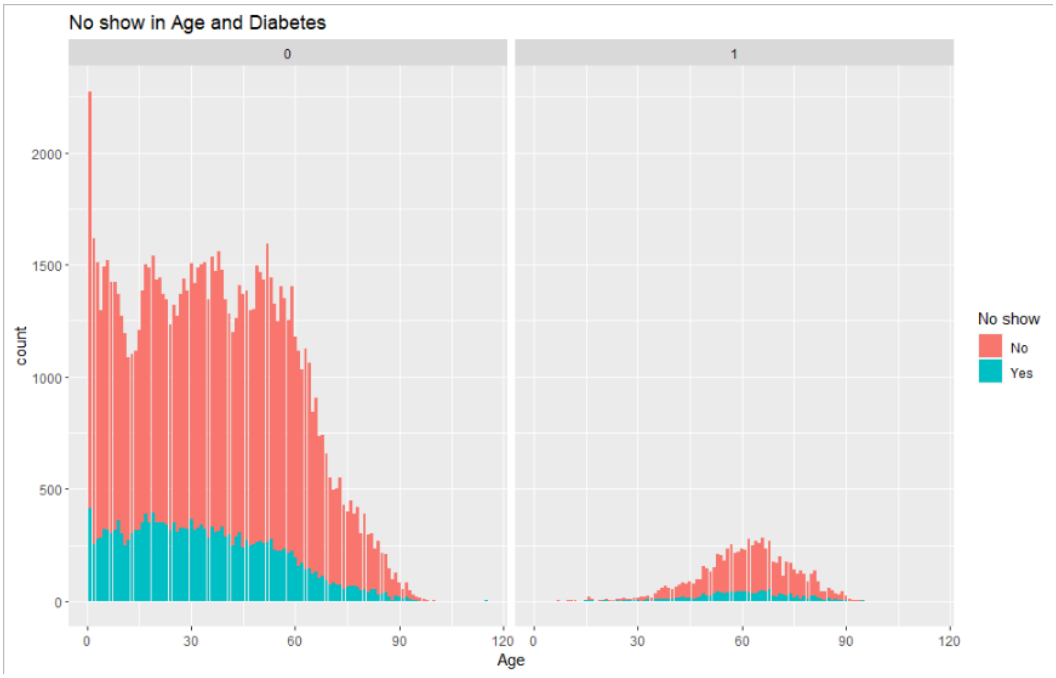
چه گروهی از بیماران No-show بیشتری دارند؟

برای اینکه بتوینم تاثیر پارامترهای مختلف را با no.show بگیریم قسمتی از کد که با هدر #---------No show details in facet --- مشخص شده به صورت ترکیب های متخلف از ۲ پارامتر مختلف به همراه no.show میگیریم و بررسی میکنیم.

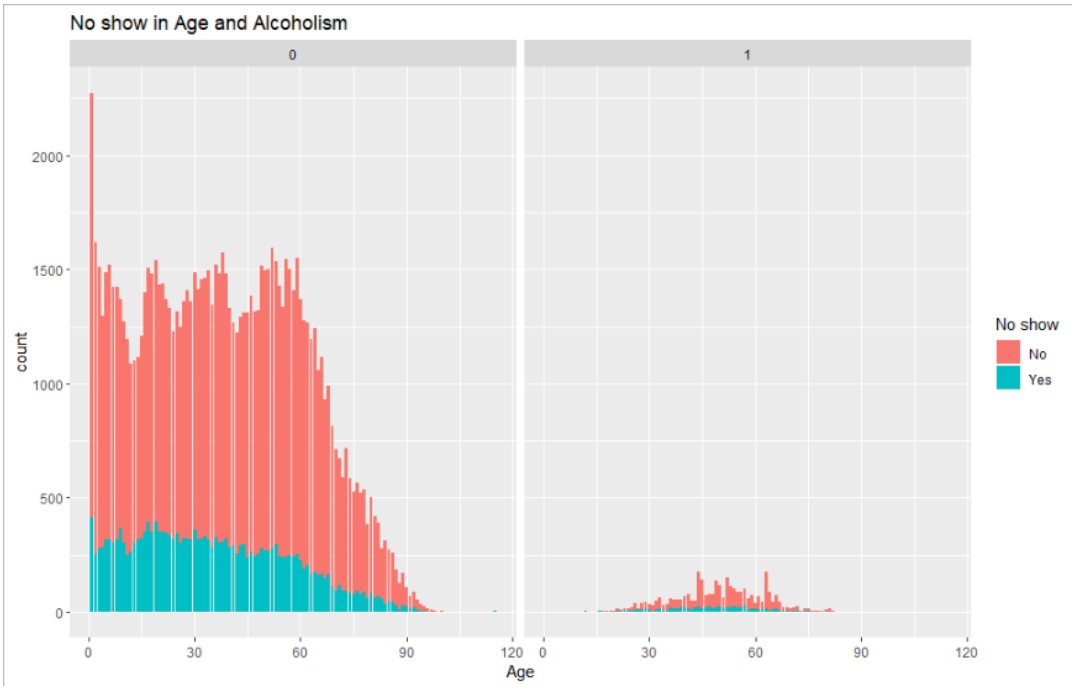


بین کسانی که SMS دریافت کردند، افرادی که کمک هزینه درمانی هم دریافت کرده بودند درصد بیشتری در قرار ملاقات حاضر شدند پس شاید اقدام مناسب برای کم کردن نرخ عدم مراجعه دربین کسانی که کمک هزینه دریافت کردند این باشد که به انها زمان قرار ملاقات را با SMS یادآوری کنیم.

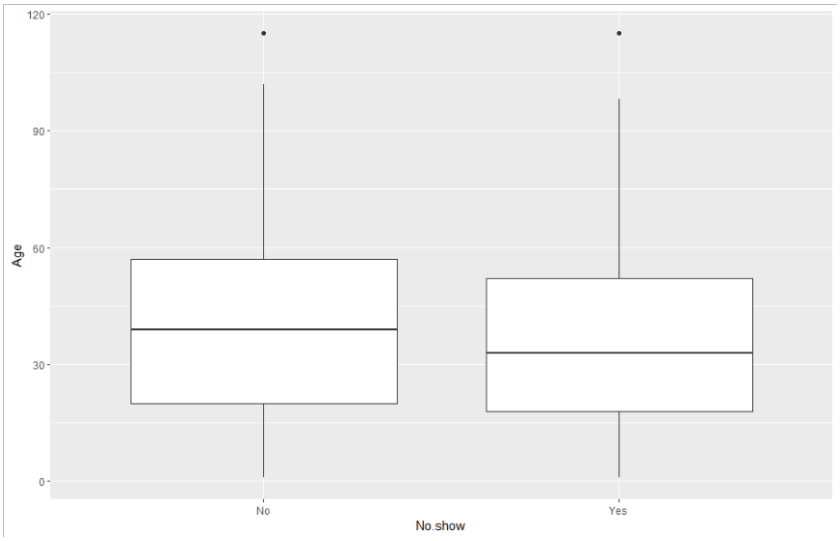
برای بررسی میزان عدم حضور بیماران در سنین مختلف و باتوجه به بیماری فشار خون نمودار بالا را رسم میکنیم. همان طو رکه انتظار داشتیم بیماری فشار خون حول افراد با سن ۶۰ یک توزیع نرمال دارد و تعداد زیادی از بیماران اطفال که نرخ عدم حضور بالایی هم داشتند زیر یکسال هستند.



برای بررسی میزان عدم حضور بیماران در سنین مختلف و باتوجه به بیماری دیابت نمودار بالا را رسم میکنیم. مطابق بیماری فشار خون این موارد در سنین ۳۰ تا ۹۰ سال و حول ۶۰ سال شیوع بیشتری دارن و در مقایسه با بیماری فشارخون نرخ عدم مراجعه کمتری دارند (قله نمودارهای سمت راستی در بیماری دیابت چیزی حدود نصف هستش)



در ادامه و بررسی بیماران الکلی مقدار عدم حضور خیلی خیلی کمتر از دو بیماری قبلی است.

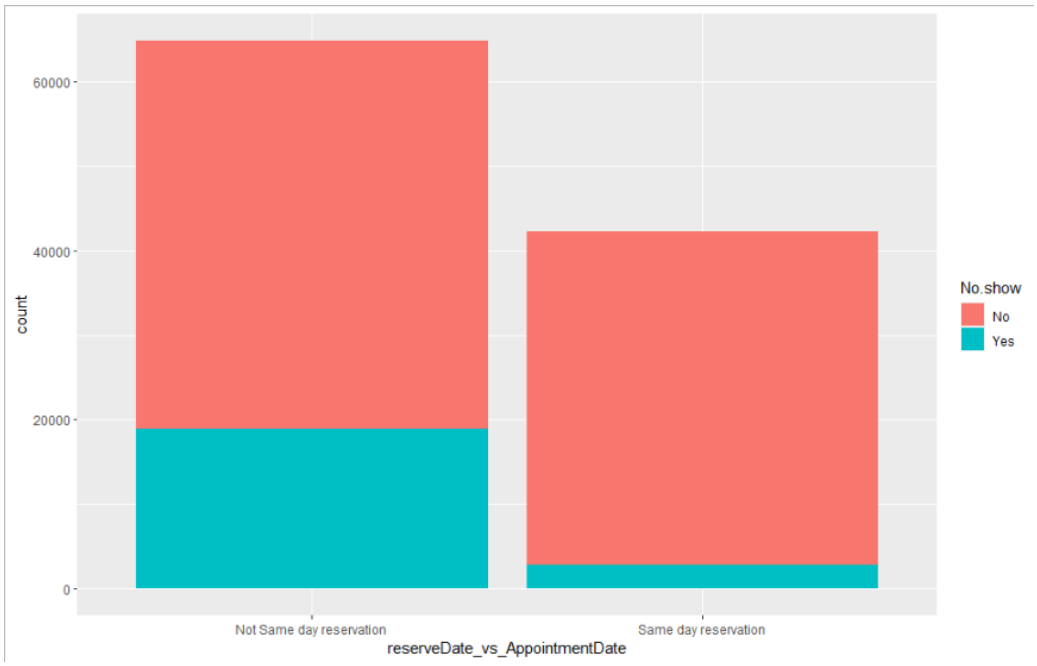


و در نهایت فقط با بررسی سن بیماران و مقایسه عدم حضور و حضور آنها، شاهد این هستیم که میانگین سنی بیمارانی که حاضر نشدند پایین تر از دسته دیگر است.

در ادامه متغیر جدیدی را که میتواند نگاه ما را نسبت به وضعیت روزرو بازتر کند را محاسبه میکنیم و به دیتاست اضافه میکنیم و تاثیر آن را با نرخ عدم مراجعه از منظرهای متفاوت بررسی میکینم.

قسمتی از کد که با هدر----------same day reservation---------# مربوط به اضافه کردن این متغیر جدید است. درواقع ایده این هست که بررسی کنیم که رزروهایی که در روز ملاقات انجام شده است چه اوضاعی در عدم حضور یا حضور داشته اند. برای این کار اختلاف بین scheduledDay و AppoinmentDay را بدست میاریم و براساس این متغیر اختلاف یک Flag تعریف میکنیم که ایا این اختلاف کمتر از یک روز (۲۴ ساعت) است مقدارSame day reservation را به رکورد تخصیصی میدهد ودر صورتی که این اختلاف بیشتر از ۲۴ ساعت باشد Not Same day reservation را تخصیص میدهد.

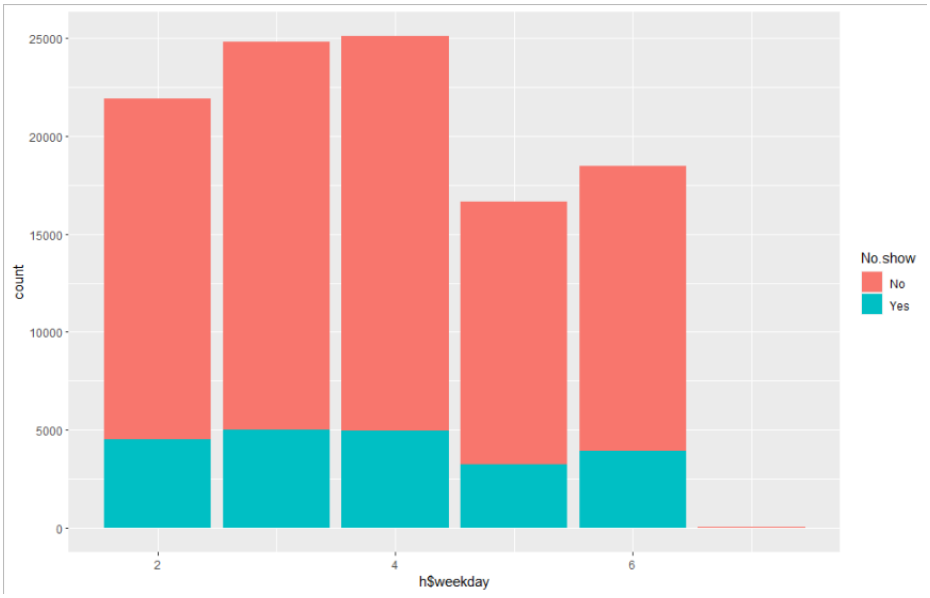
حالا میزان تاثیر این متغیر را بر عدم حضور بررسی میکنیم:



در بین افرادی که در زمان قرارملاقات حاضر نشدند تعداد بیشتری در دسته Not Same day reservation هستند . این نشون میده که برای بیمارانی که در همان روز ملاقات، رزرو را انجام نمیدهند و در روزهای قبل تر اینکار را انجام میدهند برای جلوگیری از عدم حضور انها میتوان اقداماتی را تعریف کرد. مثل ارتباط با آنها از طریق SMS یا ایمیل.

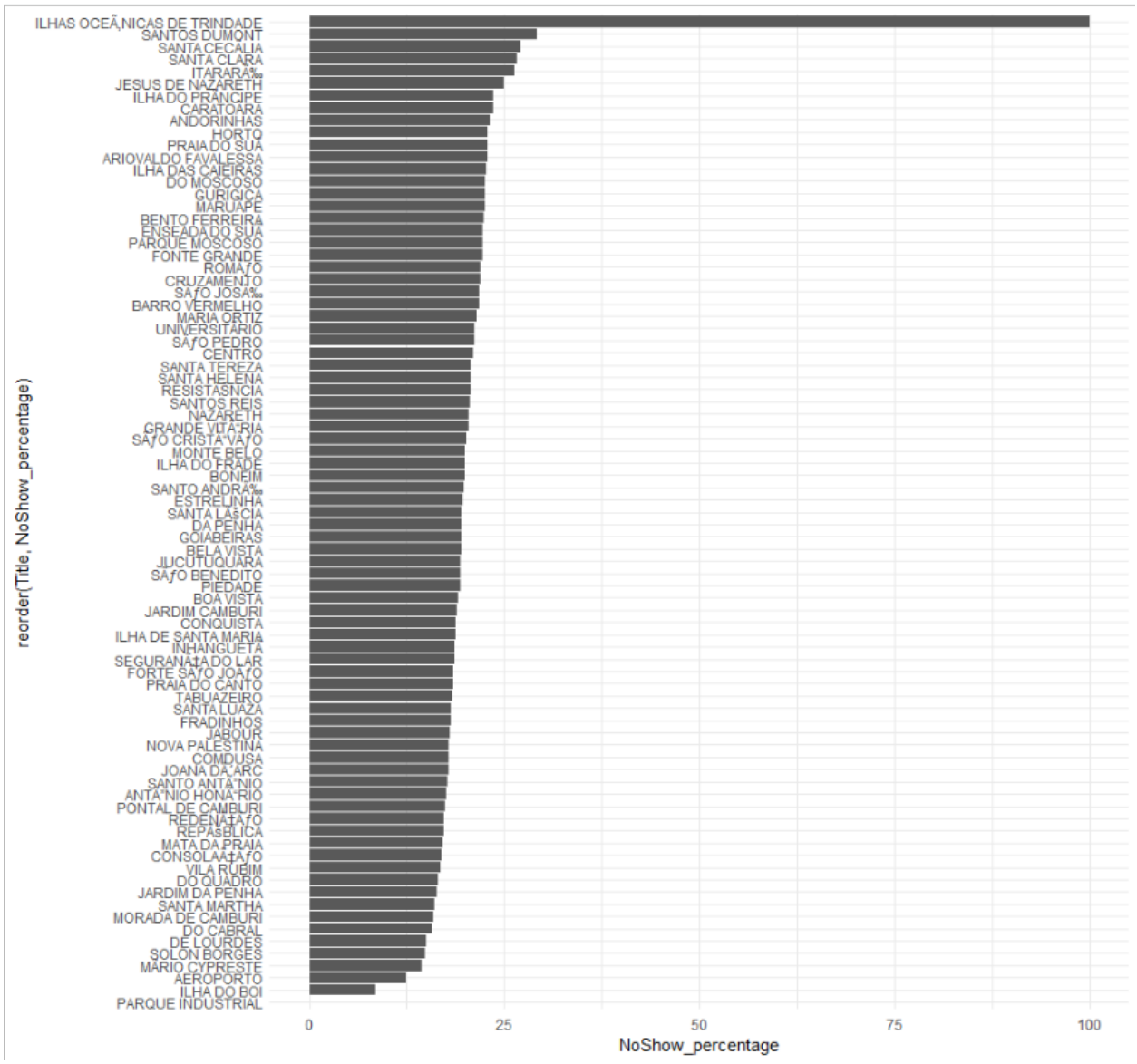
B:

آیا روز هفته در میزان show-no موثر است؟



باتوجه به اینکه در فرض های سوال آمده است که در روزهای شنبه و یکشنبه هیچ مراجعی پذیرفته نمیشود و فقط درموارد اورژانسی در روز شنبه بیمار پذیرش میشود این مطلب در نمودار هم واضح است اما درباره سهم عدم حضور، میتوان به این نکته توجه کرد که روزهای پنجشنبه و جمعه با اینکه نسبت به بقیه روزها مراجعین کمتری دارند ولی تعداد عدم مراجعه تقریبا یکسانی با بقیه روزها دارند و این نشان میدهد که سهم عدم مراجعه در روزهای ۵ شنبه و جمعه به نسبت بقیه روزها بالاتر است.

برای اینکه نرخ عدم حضور درمحله های متفاوت را داشته باشیم یک table میگیریم و بعد با محاسبه درصد عدم حضور در محله نمودار نزولی این درصدها و محله ها را رسم میکنیم، به غیر از یکی از محله ها که ۱۰۰ درصد نرخ عدم حضور داشتند (البته کلا ۲ نفر متلعق به این محله بوده اند) بقیه نرخ عدم مشابه مشابهی دارند و در یک رنج ۲۰ تا ۲۵ درصد هستند.


ILHAS OCEÃNICAS DE TRINDADE
SANTOS DUMONT
SANTA CECÃLIA
SANTA CLARA
ITARARÃ‰
JESUS DE NAZARETH
ILHA DO PRÍNCIPE
CARATOARA
ANDORINHAS
HORTO
PRAIA DO SUÁ
ARIOVALDO FAVALESSA
ILHA DAS CAIEIRAS
DO MOSCOSO
GURIGICA
MARUÍPE
BENTO FERREIRA
ENSEADA DO SUÁ
PARQUE MOSCOSO
FONTE GRANDE
ROMÃO
CRUZAMENTO
SÃO JOSÃ‰
BARRO VERMELHO
MARIA ORTIZ
UNIVERSITÁRIO
SÃO PEDRO
CENTRO
SANTA TEREZA
SANTA HELENA
RESISTÊNCIA
SANTOS REIS
NAZARETH
GRANDE VITÓRIA
SÃO CRISTÓVÃO
MONTE BELO
ILHA DO FRADE
BONFIM
SANTO ANDRÉ
ESTRELINHA
SANTA LÚCIA
DA PENHA
GOIABEIRAS
BELA VISTA
JUCUTUQUARA
SÃO BENEDITO
PIEDADE
BOA VISTA
JARDIM CAMBURI
CONQUISTA
ILHA DE SANTA MARIA
INHANGUETÁ
SEGURANÇA DO LAR
FORTE SÃO JOÃO
PRAIA DO CANTO
TABUAZEIRO
SANTA LUÍZA
FRADINHOS
JABOUR
NOVA PALESTINA
COMDUSA
JOANA D'ARC
SANTO ANTÔNIO
ANTÔNIO HONÓRIO
PONTAL DE CAMBURI
REDENÇÃO
REPÚBLICA
MATA DA PRAIA
CONSOLAÇÃO
VILA RUBIM
DO QUADRO
JARDIM DA PENHA
SANTA MARTHA
MORADA DE CAMBURI
DO CABRAL
DE LOURDES
SOLON BORGES
MÁRIO CYPRESTE
AEROPORTO
ILHA DO BOI
PARQUE INDUSTRIAL
reorder(Title, NoShow_percentage)
0
25
50
75
100
NoShow_percentage

Question #2:

آزمونهای آماری برای بررسی ادعاها و باورهای زیر تشکیل دهید:

قسمتی از کد که با هدر #--------tests----------مشخص شده است به بررسی تست های مختلف میپردازد.

A:

مسئولین سیستم سلامت معتقدند که ارسال sms برای یادآوری زمان ملاقات می تواند موجب افزایش مراجعه (کاهش no-show) بیماران شود. آیا این ادعا تایید می شود؟

تست را اینگونه تعریف میکنیم که فرض میکنیم که ارسال sms و no.show ارتباطی باهم ندارند و سعی میکنیم این فرض را با آزمون chi-square آزمون کنیم:

H0: Sending SMS to patients and No.show events are independent.

H1: Sending SMS to patients and No.show events are dependent.

Results:

Pearson's Chi-squared test with Yates' continuity correction

data: kzSMS

X-squared = 1731.8, df = 1, p-value < 2.2e-16

با توجه به نتیجه تست و اینکه مقدار p-value از ۰/۰۵ کمتر است فرض صفر را رد میکنیم و نتیجه میگیریم که این دو متغیر مستقل از یکدیگر نیستند. با این تست درمورد جهت تاثیرگذاری و موجب افزایش یا کاهش شدن، نمی توان نظری داد برای همین در ادامه یک general linear model از نوع binomial تعریف میکنیم تا با استفاده از ضریب به دست امده تاثیر مثبت یا منفی بودن را پیدا کنیم.

نتیجه مدل خطی:

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.604970	0.009957	-161.18	<2e-16	***
SMS_received1	0.643838	0.015609	41.25	<2e-16	***

باتوجه به نتیجه مدل خطی، مشاهده میشود که ۱ بودن متغیر SMS_received یعنی ارسال SMS برای بیماران به طرز محسوسی بر افزایش no.show=yes تاثیر میگذارد! یه کم نتیجه خلاف انتظاری هست ولی احتمالا نشون دهنده این هست که یک سری متغیر وابسته میانی هم باید در نظر بگیریم و اثر متغیرSMS بر روی بقیه متغیرهاست که بعد در no.show تاثیر میگذارد پس با توجه به نتیجه تست و مدل خطی موجود فقط میتونیم بگیم که این دو متغیر با هم مرتبط هستند ولی در مورد میزان اثرگذاری مستقیم باید تاثیر بقیه متغیرها را بیشتر بررسی کنیم.

B:

یکی از پزشکان بر این عقیده است که فقر مالی علت عدم مراجعه بیماران است. در این مورد با توجه به داده‌ها چه نظری دارید؟

تست را اینگونه تعریف میکنیم که فرض میکنیم که دریافت کمک هزینه درمانی و no.show ارتباطی باهم ندارند و سعی میکنیم این فرض را با آزمون chi-square آزمون کنیم:

H0: Receiving scholarship and No.show events are independent.

H1: Receiving scholarship and No.show events are dependent.

Results:

Pearson's Chi-squared test with Yates' continuity correction

data: kzScholarship

X-squared = 92.044, df = 1, p-value < 2.2e-16

با توجه به نتیجه تست و اینکه مقدار p-value از ۰/۰۵ کمتر است فرض صفر را رد میکنیم و نتیجه میگیریم که این دو متغیر مستقل از یکدیگر نیستند. با این تست درمورد جهت تاثیرگذاری و موجب افزایش یا کاهش شدن، نمی توان نظری داد برای همین در ادامه یک general linear model از نوع binomial تعریف میکنیم تا با استفاده از ضریب به دست امده تاثیر مثبت یا منفی بودن را پیدا کنیم.

نتیجه مدل خطی:

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.23594	0.01159	-192.92	<2e-16	***
No.showYes	0.23008	0.02399	9.59	<2e-16	***

باتوجه به نتیجه مدل خطی، مشاهده میشود که ۱ بودن متغیر Scholarship یعنی دریافت کمک هزینه توسط بیماران به طرز محسوسی بر افزایش no.show=yes تاثیر میگذارد. باتوجه به نتیجه موجود میتوان نظر پزشک را رد کرد و شاید اینگونه نتیجه را تفسیر کرد که اتفاقا در اختیار قرار دادن کمک هزینه مالی به بیماران بدون آگاهی سازی بیماران باعث میشود که بیماران تعهدی برای حضور در وقت ملاقات حس نکنند!

C:

شورای سیاست گذاری یکی از بیمارستانها عقیده دارد که نوع بیماری در عدم مراجعه موثر است و در این میان بیماران الکلی بیشترین نرخ عدم مراجعه را دارند.

ابتدا تست را روی بیماران با اعتیاد الکل شروع میکنیم و بعد بقیه بیماری ها را بررسی میکنیم.

تست را اینگونه تعریف میکنیم که فرض میکنیم که ابتلا به اعتیاد به الکل و no.show ارتباطی باهم ندارند و سعی میکنیم این فرض را با آزمون chi-square آزمون کنیم:

H0: Alchoholism and No.show events are independent.

H1: Alchoholism and No.show events are dependent.

Results:

Pearson's Chi-squared test with Yates' continuity

correction

data: kzAlcoholism

X-squared = 0.021664, df = 1, p-value = 0.883

با توجه به نتیجه تست و اینکه مقدار p-value از ۰/۰۵ بیشتر است فرض صفر را میپذیریم و نتیجه میگیریم که این دو متغیر مستقل از یکدیگر هستند و مبتلا بودن به اعتیاد به الکل ربطی به پدیدهno.show ندارد.

مشابه آزمون الکل، برای بیماری های دیابت، فشار خون و همچنین معلولیت هم ازمون های مناسب را اجرا میکنیم تا بتونیم در مورد ارتباط بیماری های مختلف و پدیده no.show نظر دهیم. به طور خلاصه مقادیر p-value در تست ها در جدول زیر امده است:

بیماری	p-value of Chi-square test	Result
Hypertension	p-value < 2.2e-16	**H0 is rejected and they are dependent**
Diabetes	p-value = 2.043e-07	**H0 is rejected and they are dependent**
Handicap	p-value = 0.1172	**H0 is accepted and they are independent**

Question #3:

با توجه به نمودارهایی که رسم کرده اید، چه دلایلی را برای عدم مراجعه بیماران حدس می زنید؟

باتوجه به نمودارها و تست‌های قبل به نظر میرسد که سن پایین و تخصیصی دادن قرار ملاقات در روزهای بعد از رزرو، و قرارملاقات در روزهای ۵ شنبه و جمعه درصد بالاتری از no.show را به خودشون اختصاص دادند.

طولانی شدن صف و اینکه قرارملاقات در روز رزرو انجام نمیشود یکی ازدلایل پررنگ دیگری است که نرخ عدم حضور را تحت تاثیر قرار میدهد.

در کل به نظر میرسد که با داشتن دیتای بیشتر و attribute های بیشتر بتونیم بهتر این پدیده no.show را بررسی کنیم.

Question #4:

با توجه به بررسی داده ها پیشنهاداتی را در جهت کاهش هزینه های عدم مراجعه به شورای سیاستگذاری مراکز درمانی ارایه دهید.

باتوجه به نموداها و تست‌های قبل به نظر میرسد که درصد بالایی از عدم مراجعه ها مربوط به بیماران اطفال است و همچنین مراجعینی که در همان روز قرار ملاقات رزرو را انجام نداده اند هم فراوانی بیشتری دارند، لذا اقدام مناسبی که برای این دسته از بیماران به نظر میرسد این است که برا اینکه بتوان نرخ مراجعه را برای این دسته از بیماران کم کرد این است که بتوان در همان روز رزرو، به آنها قرار ملاقات تخصیص داد که برای اینکار توصیه میشود که صف مربوط به بیماران اطفال را جدا کنیم.

اقدام مناسب دیگه ای که میتوان پیشنهاد داد این است که باتوجه به نرخ عدم حضور بالای بیماران اطفال و سیاست کلی مرکز درمانی درعدم پذیرش در روهای تعطیل، قسمتی از منابع مرکز درمانی را برای روزهای تعطیل فعال کنیم و فقط مراجعین اطفال را به این روزها اختصاص دهیم با این اقدام احتمالا خانواده این بیماران اطفال در روزهای تعطیل بهتر میتوانند در مرکز درمانی حاضر شوند. همچنین برای آن دسته از بیمارانی که کمک هزینه درمانی دریافت کرده اند SMS قرار ملاقات را ارسال کنیم.

یکی از اقدامات میتواند این باشد که بودجه SMS را به طور خاص به افرادی که در ۳ محله که نرخ عدم مراجعه بالاتر دارند اختصاص دهیم.

برای جلوگیری از هدررفت منابع میتوان یک تابع احتمالی با توجه به دتیای موجود حساب کرد که یک احتمال بین ۰ و ۱ به هر appointment id اختصاص داد و بعد با توجه به بالا بودن این احتمال (بالا بودن از یک تریشلد مشخص و با توجه به بودجه موجود) برنامه آگاه کردن مراجعین را از طریق کانال های ارتباطی مثل SMS و ایمیل و تماس انجام داد.