



HACETTEPE ÜNİVERSİTESİ
İSTATİSTİK BÖLÜMÜ

İST 347 – İSTATİSTİKSEL ANALİZ
ARA SINAV ÖDEVİ

2200329056 Beyza Senanur AKTAŞ

2200329046 Hilal KAYA

Prof. Dr. Onur TOKA

Aralık, 2024

İÇİNDEKİLER

1. GİRİŞ	4
2. ANALİZ	5
2.1 Göreli Frekans Hesaplama.....	5
mean () Fonksiyonu:	5
2.2 Medyan İçin Güven Aralığı Hesaplama	6
wilcox.test () Fonksiyonu:	6
2.3 Örneklem Oranını Test Etme	8
prop.test () Fonksiyonu:	8
2.4 Bir Oran İçin Güven Aralığı Oluşturma	9
2.5 Korelasyonun Anlamlılığını Test Etme	11
cor.test () Fonksiyonu.....	11
2.6 ggplot Grafiklerini Özelleştirme.....	12
2.7 Bar Grafiği Renklendirme	14
geom.bar () Fonksiyonu:.....	14
2.8 Grafiklere Yatay ve Dikey Çizgi Ekleme	17
geom_hline () ve geom_line () Fonksiyonu:.....	17
2.9 Faktör Düzeyleri İçin Boxplot Grafipleri	19
2.10 Histograma Yoğunluk Tahmini Ekleme.....	20
geom_density() Fonksiyonu:	20
2.11 Normal (Q-Q) Grafiği Oluşturma.....	22
2.12 Değişkenleri Farklı Renklerle Görselleştirme	23
case when() Fonksiyonu	23
mutate() Fonksiyonu:	23
2.13 Basit Doğrusal Regresyon Modeli.....	24
lm() Fonksiyonu:.....	24
2.14 Çoklu Doğrusal Regresyon.....	25
2.15 Summary.....	26
summary() Fonksiyonu:	26
2.16 Dönüştürülmüş Veri Regresyonu.....	29
log() Fonksiyonu:.....	29

2.17 Regresyon Modelinin Artık Değerlerini Görselleştirme	31
augment() Fonksiyonu:	31
2.18 Etkili Gözlemleri Tanımlama.....	33
influence.measures() Fonksiyonu:	33
2.19 One-way ANOVA	34
oneway.test () Fonksiyonu:	34
3.SONUÇ	35
4. KAYNAKÇA	36

1.GİRİŞ

Veri Kümesi Hakkında:

Bu çalışmada, bireylerin gelir düzeyleri, harcama alışkanlıkları ve demografik özellikleri incelenmiştir. 2000 gözlemden veriler elde edilmiştir.

Veri Kümesi Değişkenleri:

- ★ **Income:** Yıllık gelir (para birimi bazında).
- ★ **Age:** Bireyin yaşı.
- ★ **Dependents:** Bakımından sorumlu olduğu kişi sayısı.
- ★ **Occupation:** Çalışma durumu (örneğin, öğrenci, emekli).
- ★ **City Tier:** Bireyin yaşadığı şehrin gelişmişlik düzeyi (1, 2 veya 3).
- ★ **Rent:** Aylık kira harcaması.
- ★ **Loan Repayment:** Aylık kredi geri ödeme tutarı.
- ★ **Groceries:** Aylık market harcaması.
- ★ **Transport:** Ulaşım giderleri.
- ★ **Eating Out:** Dışarıda yemek için yapılan harcamalar.
- ★ **Entertainment:** Eğlence için yapılan harcamalar.
- ★ **Utilities:** Elektrik, su gibi kamu hizmeti giderleri.
- ★ **Healthcare:** Sağlık hizmetlerine yapılan harcamalar.
- ★ **Education:** Eğitim için yapılan harcamalar.
- ★ **Miscellaneous:** Çeşitli ek giderler.

Amaç:

Bu rapor, gelir ile harcama alışkanlıkları arasındaki ilişkiyi analiz etmek, bireylerin finansal davranışlarını anlamak ve demografik faktörlerin etkisini ortaya koymak için analiz yapmak amaçlamaktadır.

Bunun için keşifsel veri analizi, hipotez testleri ve regresyon modelleri kullanılmıştır.

2. ANALİZ

2.1 Göreli Frekans Hesaplama

mean () Fonksiyonu:

Fonksiyon veri setindeki sayısal değerlerin ortalamasını bulunmasına yardımcı olur. Değişken ortalamasının belli bir değer üzerinde olup olmadığı analiz edilebilir.

- Yaş değişkeni gruplandırılarak her bir grubun ortalama geliri hesaplanır ve bunun üzerinden karşılaştırma yapılır.

```
> cat("18-25 yaş grubu ortalama gelir: ", mean_income_18_25 /12 , "\n")
18-25 yaş grubu ortalama gelir: 3756.006
> cat("26-35 yaş grubu ortalama gelir: ", mean_income_26_35 /12 , "\n")
26-35 yaş grubu ortalama gelir: 3390.242
> cat("36-50 yaş grubu ortalama gelir: ", mean_income_36_50 /12 , "\n")
36-50 yaş grubu ortalama gelir: 3407.473
> cat("51-65 yaş grubu ortalama gelir: ", mean_income_51_65 /12 , "\n")
51-65 yaş grubu ortalama gelir: 3613.057
```

Yaş Grubu	Ortalama Gelir (Aylık)
18-25	3756.006
26-35	3390.242
36-50	3407.473
51-65	3613.057

- Her yaş grubunda ortalama gelirler birbirine yakın seyretmektedir. Ancak en yüksek aylık gelirin 18-25 grubunda olduğu gözlemlenir. Buna karşın en düşük aylık gelir 26-35 yaş grubunda görülür.

2.2 Medyan İçin Güven Aralığı Hesaplama

wilcox.test () Fonksiyonu:

Fonksiyon Wilcoxon testini seçilen değişkene uygular. Wilcoxon testi, iki bağımlı örneklem arasında ortanca farkını test etmek için kullanılan parametrik olmayan bir istatistiksel testtir.

Tek Örneklem Durumu: Tek örneklemin medyanının belirlenen değerden farklı olup olmadığını test etmek için kullanılır. Örneğin, bir grup bireyin gelir seviyelerinin medyanı ile 0'a olan farkı test etmek.

Bağımlı İki Örneklem Durumu: Aynı gruptan alınan iki farklı veri seti için, bu iki veri seti arasındaki farkı test etmek için kullanılır. Örneğin, bir grup öğrencinin sınavdan önce ve sonra aldıkları puanlar arasındaki fark test edilir.

- Eğitim değişkeni için güven aralığı ve medyan hesaplanır.

```
#Aylık eğitim harcaması için güven aralığı ve medyan
> wilcox.test(education/12, conf.int = TRUE)

      wilcoxon signed rank test with continuity correction

data:  education/12
V = 1247410, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 0
95 percent confidence interval:
 204.9662 224.0484
sample estimates:
(pseudo)median
 214.2766
```

%95 güven düzeyinde,

Medyan = 214.28

Güven Aralığı = [204.97, 224.05]

p-değeri = 2.2e-16

- Bu sonuçlara göre, aylık eğitim harcamalarının medyanı 214.28'dir ve medyanın sıfırdan anlamlı bir şekilde farklı olduğu görülmektedir. p-değeri çok düşük olduğu için de bu farkın olduğu söylenebilir. Medyan güven aralığındaki değerlerin içinde olduğundan bir güven verir.
- Burada da güven aralığı 2.1 başlığında hazırlanan yaş grupları için hesaplanır.

```
> #Yaş gruplarına göre aylık eğitim harcaması için güven aralığı ve medyan
> wilcox.test(education[freq_18_25], conf.int = TRUE)

      wilcoxon signed rank test with continuity correction

data:  education[freq_18_25]
V = 31878, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 0
95 percent confidence interval:
 2273.121 2796.757
sample estimates:
(pseudo)median
 2519.954

> wilcox.test(education[freq_26_35], conf.int = TRUE)

      wilcoxon signed rank test with continuity correction

data:  education[freq_26_35]
V = 42486, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 0
95 percent confidence interval:
 2321.259 2908.410
sample estimates:
(pseudo)median
 2606.177

> wilcox.test(education[freq_36_50], conf.int = TRUE)

      wilcoxon signed rank test with continuity correction

data:  education[freq_36_50]
V = 115921, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 0
95 percent confidence interval:
 2341.836 2737.970
sample estimates:
(pseudo)median
 2533.146

> wilcox.test(education[freq_51_65], conf.int = TRUE)

      wilcoxon signed rank test with continuity correction

data:  education[freq_51_65]
V = 108811, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 0
95 percent confidence interval:
 2387.364 2825.125
sample estimates:
(pseudo)median
 2598.704
```

Yaş Grubu	Medyan	95% Güven Aralığı
18-25 yaş	2519.95	2273.12 ile 2796.76
26-35 yaş	2606.18	2321.26 ile 2908.41
36-50 yaş	2533.15	2341.84 ile 2737.97
51-65 yaş	2598.70	2387.36 ile 2825.13

- Yaş gruplarının eğitim harcamaları medyan olarak benzerdir, ancak 26-35 yaş grubu biraz daha yüksek harcamaya sahiptir.
- Eğitim harcamalarındaki bu benzerlik, yaş farkına rağmen eğitimin genel bir öncelik olduğunu gösterebilir.

2.3 Örneklem Oranını Test Etme

prop.test () Fonksiyonu:

Bu fonksiyon, bir örnekteki başarı oranının belirli bir değere karşı test edilmesini sağlar. Bu test, özellikle binom dağılımına dayalı testler için kullanılır.

`prop.test(x, n, p, alternative = "greater")`

x: Başarı sayısı.

n: Toplam örneklem büyüklüğü.

p: Sıfır hipotezine göre beklenen başarı oranı

- `alternative`: Hipotez türü:
- `"two.sided"`: İki yönlü test (Oranın beklenen değere eşit olup olmadığı test edilir).
- `"greater"`: Tek yönlü test (Oranın beklenen değerden büyük olup olmadığı test edilir).
- `"less"`: Tek yönlü test (Oranın beklenen değerden küçük olup olmadığı test edilir).

➤ Tier_1 şehrinde yaşayanların oranının %50'den büyük olup olmadığına bakılır.

```
> x <- sum(city == "Tier_1")
> n <- length(city)
> p <- 0.5
> test_result <- prop.test(x, n, p , alternative = "greater")
> test_result
```

1-sample proportions test with continuity correction

```
data: x out of n, null probability p
X-squared = 388.08, df = 1, p-value = 1
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.263056 1.000000
sample estimates:
      p
0.2795
```

X-squared = 388.08
p-değeri = 1
Alternatif Hipotez: Gerçek oran, %50'den büyüktür.
Güven Aralığı (95%): [0.2631, 1.0000]
Örnekleme Oranı (p): 0.2795

- "Tier_1" şehirlerinin oranının %50'den büyük olduğu hipotezi p-değeri 1 olduğu için reddedilemez. Bu, "Tier_1" şehirlerinin oranının %50'den fazla olduğuna dair istatistiksel olarak anlamlı bir kanıt olmadığı anlamına gelir.

2.4 Bir Oran İçin Güven Aralığı Oluşturma

Analizin bu kısmında yine prop.test () fonksiyonu kullanılıyor.

- Kredi ödemesi olmayanların popülasyondaki oranının %50 olup olmadığına bakılır.

```

> #ORAN TESTİ GÜVEN ARALIĞI
> x <- sum(loan == 0)
> n <- length(loan)
> confidence_interval <- prop.test(x, n)
> confidence_interval

1-sample proportions test with continuity correction

data: x out of n, null probability 0.5
X-squared = 109.98, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.5957424 0.6388018
sample estimates:
      p
0.6175

```

H₀: Popülasyondaki oran $p = 0.5$

H_s: Popülasyondaki oran $p \neq 0.5$

- p değeri $= 2.2e-16 < 0$ olduğu için H_0 reddedilir. Oranın %50'den farklı olduğu söylenebilir.
- $p = 0.6175$ Kredi ödemesi olmayan bireylerin örneklemdaki oranı = %62
- Güven Aralığı (95%): Oranın %59.57 ile %63.88 arasında olduğu tahmin edilmektedir.

Sonuçlara göre kredi ödemesi olmayan bireylerin popülasyondaki oranının %50'ye eşit değildir.

Örneklemdaki oran %61,75'tir ve bu oran %50'den anlamlı derecede büyüktür.

Güven Aralığı, popülasyon oranının %59,57 ile %63.88 arasında olduğunu desteklemektedir. %50 bu aralığa dahil edilmediği için, popülasyon oranının %50'den farklı olduğu söylenebilir.

2.5 Korelasyonun Anlamlılığını Test Etme

`cor.test ()` Fonksiyonu

İki değişken arasındaki korelasyonun anlamlı olup olmadığını test eder. Korelasyonun p-değeri ve güven aralığını hesaplar. Veri normal dağılıyorsa Pearson, normal dağılmıyorsa Spearman yöntemleri kullanılır.

- Market ve dışarıda yeme harcamaları için korelasyon analizi.

```
#KORELASYON ANLAMLILIĞI TEST ETME
> shapiro.test(groceries/12)

      Shapiro-Wilk normality test

data:  groceries/12
W = 0.61413, p-value < 2.2e-16

> shapiro.test(eating_out/12)

      Shapiro-Wilk normality test

data:  eating_out/12
W = 0.64731, p-value < 2.2e-16

> correlation <- cor.test(groceries, eating_out, method = "spearman")
> correlation

      Spearman's rank correlation rho

data:  groceries and eating_out
S = 88185444, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9338609
```

Shapiro-Wilk Normallik Testi

Market Harcamaları

- $W = 0.614$,
- $p\text{-değeri} = 2.2e-16$

Veriler normal dağılım göstermemektedir.

Dışarıda Yemek Harcamaları

- $W = 0.647$
- $p\text{-değeri} = 2.2e-16$

Veriler normal dağılım göstermemektedir.

Spearman Korelasyon Testi

- Korelasyon Katsayısı (ρ): 0.934
- $p\text{-değeri} = 2.2e-16$

Market ve dışarıda yemek harcamaları arasında pozitif ve güçlü bir ilişki bulunmaktadır. Bu ilişkinin katsayı değeri %93,4'tür.

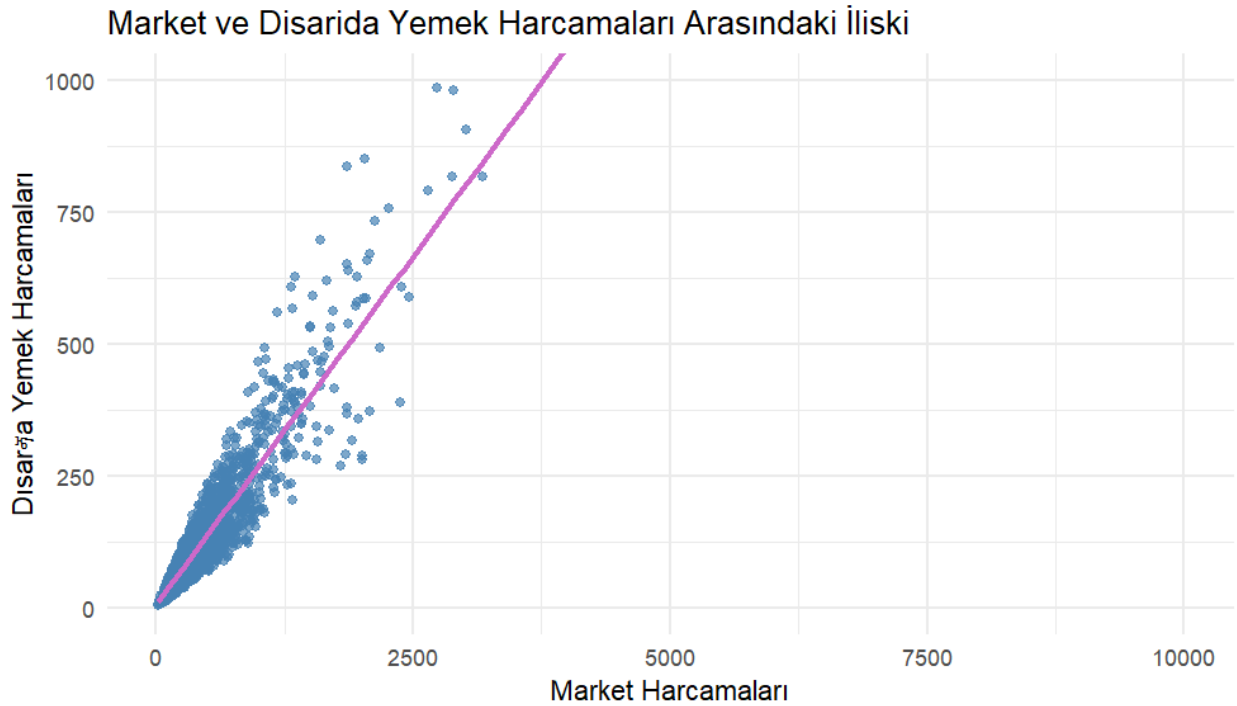
2.6 ggplot Grafiklerini Özelleştirme

ggplot2, istatistiksel veya veri görselleştirmeleri oluşturmak için kullanılan bir R paketidir. Çoğu diğer grafikleme paketinden farklı olarak, ggplot2 bir altyapı dilbilgisine sahiptir. Bu dilbilgisi, "**Grammar of Graphics**" (Wilkinson, 2005) üzerine inşa edilmiştir ve bağımsız bileşenleri bir araya getirerek grafik oluşturulmasına olanak tanır.

theme () fonksiyonu, ggplot grafiklerinde kullanılan stil ve formatları ayarlamak için kullanılır. ggplot2 paketinde, grafikleri hızla özelleştirebilecek önceden tanımlanmış birçok tema bulunmaktadır. Bu temalar, grafiklerin renkleri, yazı tipleri, eksenler, arka planlar ve diğer stil özellikleriyle ilgilidir.

theme_bw(): Beyaz arka plan ve siyah eksenler.
theme_dark(): Koyu arka plan ve açık renkler.
theme_classic(): Klasik ve sade bir stil.
theme_gray(): Varsayılan gri tema.
theme_linedraw(): Çizgisel bir stil.
theme_light(): Açık arka plan ve yumuşak renkler.
theme_minimal(): Minimalist bir stil, az detayla sadeleştirilmiş.
theme_test(): Deneme amaçlı tema.
theme_void(): Hiçbir detay veya eksen içermez, tamamen boş bir grafik.

- Market ve dışarıda yemek harcamaları arasındaki ilişki bu bölümde dağılım grafiği ve regresyon doğrusu ile görselleştirilmiştir.



- Noktaların çoğu düşük harcama düzeylerinde yoğunlaşmış, bu da bireylerin genelde hem market hem de yemek harcamalarının düşük olduğunu göstermektedir.
- Regresyon çizgisi, iki değişken arasında pozitif bir ilişki olduğunu ifade ediyor: Market harcamaları arttıkça dışarıda yemek harcamaları da artar.
- Bu ilişki, Spearman korelasyon katsayısı ($\rho = 0.93$) ile de desteklenir.

Sonuç olarak, bu iki harcama türü birbirini tamamlayan bir ilişkiye sahip denilebilir.

2.7 Bar Grafiği Renklendirme

geom.bar () Fonksiyonu: Kategorik değişkenleri görselleştirmek için çubuk grafikleri (bar plots) oluşturmada kullanılır. Bu fonksiyon, bir veri setindeki kategorilere göre frekansları veya belirtilen değerlerin toplamını görselleştirir.

dplyr Kütüphanesi:

- Veri manipülasyonu için temel bir kütüphane.
- Fonksiyonlar: group_by, summarise, mutate, %>% (pipe operatörü).

forcats Kütüphanesi:

- Faktörlerle çalışmayı kolaylaştırır.
- Fonksiyon: fct_inorder.

ggplot2 Kütüphanesi:

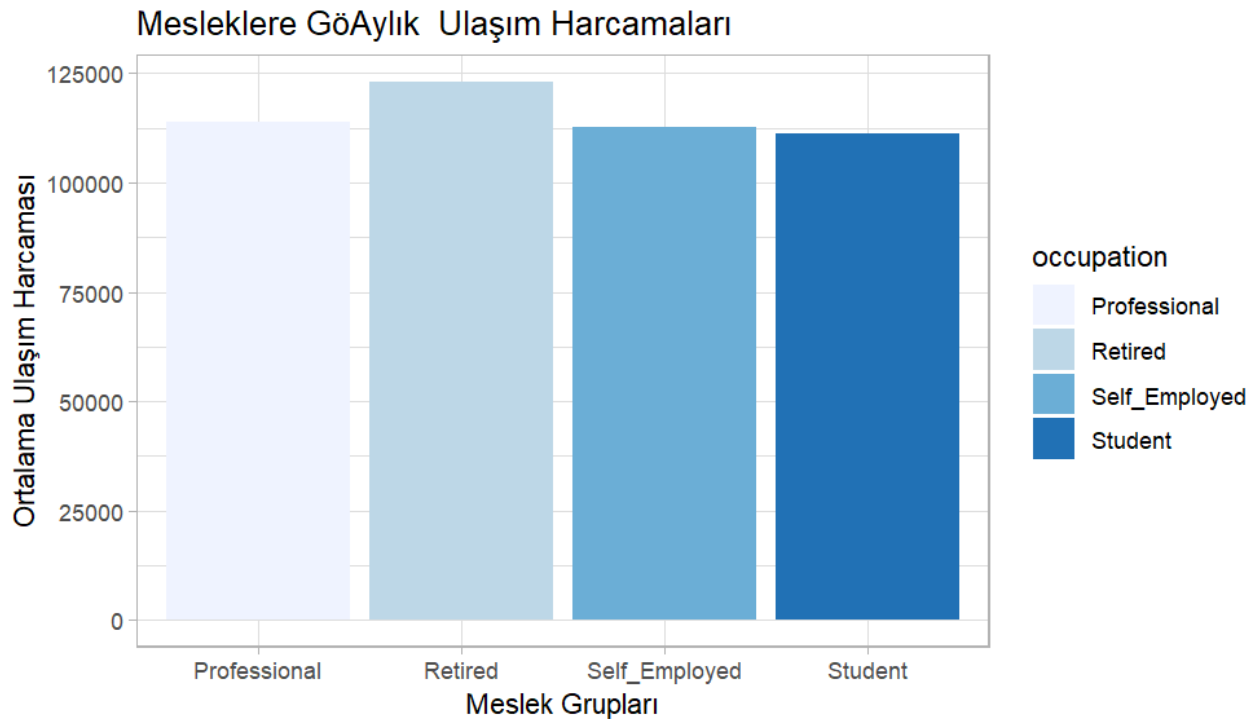
- Gelişmiş veri görselleştirme araçlarını içerir.
- Fonksiyonlar: ggplot, geom_bar, labs, scale_fill_brewer.

➤ Meslek gruplarına göre ulaşım harcamaları bar grafikleri çizilir.

```

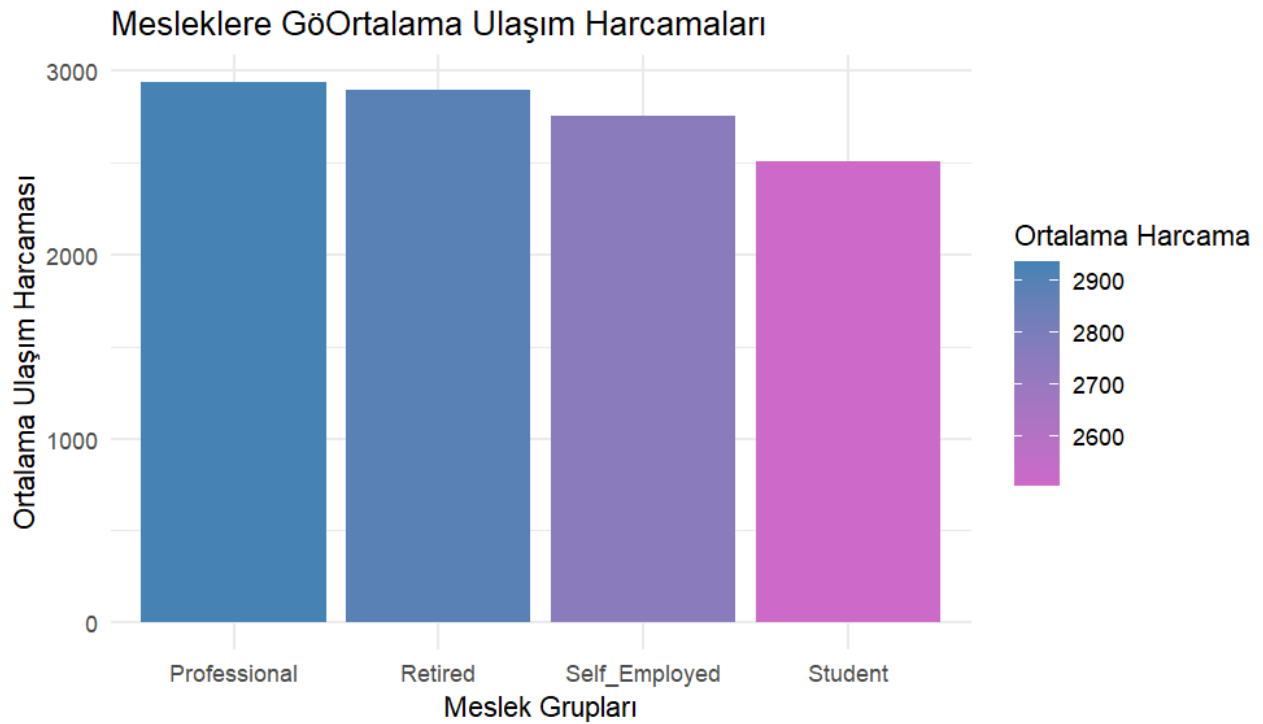
> library(dplyr)
> library(forcats)
> library(ggplot2)
> new_data <- data.frame(occupation, transport) %>%
+   group_by(occupation) %>%
+   summarise(mean_entertainment = mean(transport/12, na.rm = TRUE)) %>%
+   mutate(occupation = fct_inorder(occupation))
> # Grafik çizimi
> ggplot(data = new_data, aes(x = occupation, y = transport/12, fill = occupation)) +
+   geom_bar(stat = "identity") +
+   labs(title = "Mesleklere Göre Aylık Ulaşım Harcamaları",
+        x = "Meslek Grupları",
+        y = "Ortalama Ulaşım Harcaması")
+   ) +
+   scale_fill_brewer(palette = "Blues") + theme_light()
> ggplot(data, aes(x = occupation, y = transport, fill = ..y..)) +
+   geom_bar(stat = "summary", fun = "mean") +
+   labs(title = "Mesleklere Göre Ortalama Ulaşım Harcamaları",
+        x = "Meslek Grupları",
+        y = "Ortalama Ulaşım Harcaması",
+        fill = "Ortalama Harcama") +
+   scale_fill_gradient(low = "orchid3", high = "steelblue") +
+   theme_minimal()

```



➤ Bu grafikte meslek gruplarına (occupation) göre aylık ulaşım harcamalarının ortalamaları gösterilmektedir.

- "Retired" (Emekli) ve "Student" (Öğrenci) gruplarının ulaşımına daha fazla harcama yaptığı görülür.
- "Professional" (Profesyonel) grubunun ulaşım harcamaları en düşük seviyede gözlemlenir.
- Tüm gruplar arasında ciddi farklar gözlenmese de emekli grubunun harcamaları diğerlerine göre daha yüksektir.



Bu grafikte de meslek gruplarına göre ulaşım yapılan ortalama harcamalar gösterilmiştir. Ancak bu grafikte renk kodlaması harcama miktarlarına göre bir gradyan kullanılarak verilmiştir.

- Tüm gruplar arasında harcamalar 2600 ile 2900 birim arasında değişmektedir.
- "Professional" ve "Retired" grupları birbirine çok benzer harcama seviyelerine sahiptir.
- "Self_Employed" ve "Student" grupları da benzer şekilde gruplanmış, ancak biraz daha düşük harcamalara sahiptir.

Bu grafik, daha önceki grafiğe kıyasla daha detaylı renk ayrımı sunarak, gruplar arasındaki küçük farklılıkları görselleştirmektedir. Ancak genel olarak gruplar arasında çok büyük bir fark bulunmadığı söylenebilir.

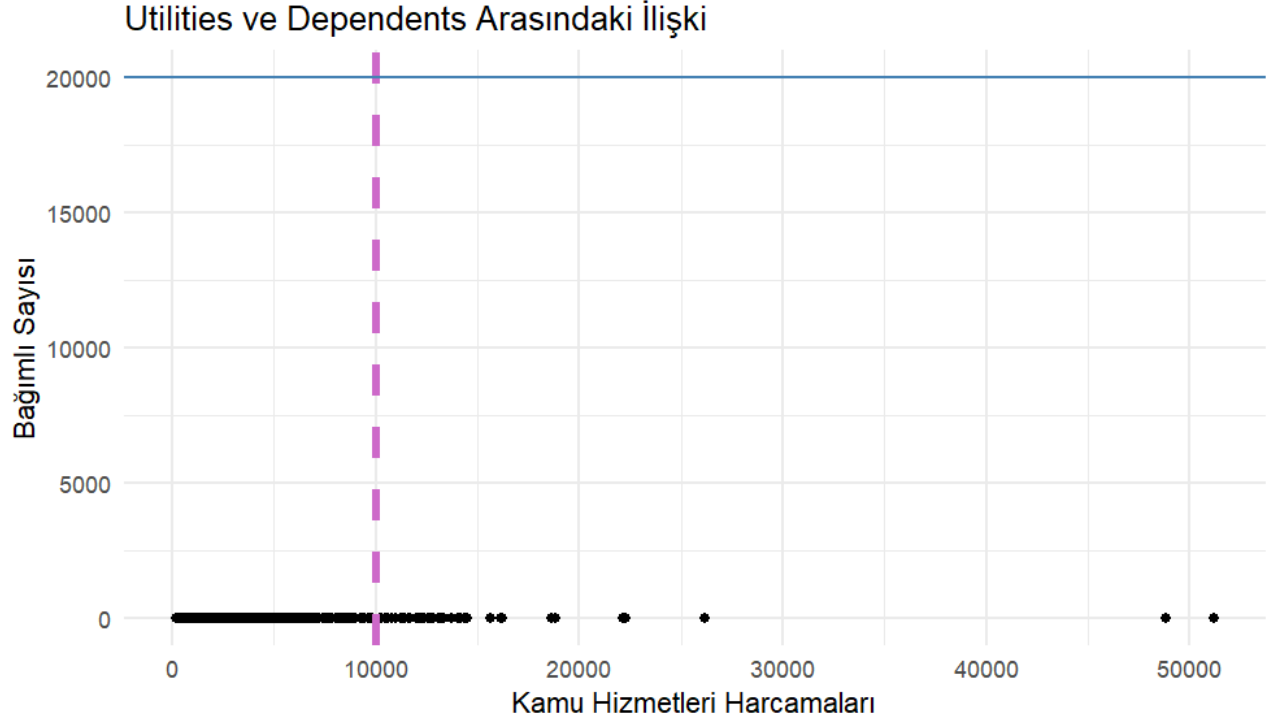
2.8 Grafiklere Yatay ve Dikey Çizgi Ekleme

geom_hline () ve geom_vline () Fonksiyonu:

Bir veri görselleştirmesi yaparken belirli bir değeri vurgulamak için grafiklere yatay (geom_hline) veya dikey (geom_vline) çizgiler eklemeyi öğrenmek. Örneğin:

- Ortalama bir değeri göstermek,
- Bir eşik değerini belirtmek,
- İstatistiksel standart sapma aralıklarını işaretlemek.

```
> ggplot(data) +
+   aes(x = utilities, y = dependents) +
+   geom_point() +
+   geom_vline(
+     xintercept = 10000,
+     color = "orchid3",
+     linetype = "dashed",
+     size = 1.5
+   ) +
+   geom_hline(
+     yintercept = 20000,
+     color = "steelblue"
+   ) +
+   labs(
+     title = "Utilities ve Dependents Arasındaki İlişki",
+     x = "Kamu Hizmetleri Harcamaları",
+     y = "Bağımlı Sayısı"
+   ) +
+   theme_minimal()
```



- Kişinin sorumlu olduğu kişi sayısı ve kamu hizmeti harcamaları arasındaki ilişkiye bakılır.
- Veriler, çoğunlukla 10.000 birim altındaki kamu hizmetleri harcamaları ve 1.000 kişi altındaki bağımlı sayılarında yoğunlaşmıştır. Bu durum, düşük harcama ve küçük aile yapılarını işaret eder.
- Dikey çizgi yüksek harcamaları sınırlandırırken, yatay çizgi büyük bağımlı grupların veri setinde yer almadığını göstermektedir. Kamu hizmetleri harcamalarının bağımlı sayısından çok gelir düzeyi ve bireysel tercihlerle belirlendiği anlaşılır.
- Bu grafik, kamu hizmetleri harcamalarının artırılmasına yönelik teşvikler ve daha büyük aile yapılarına ilişkin politikaların geliştirilmesi gerektiğine işaret eder.

2.9 Faktör Düzeyleri İçin Boxplot Grafileri

Boxplot (Kutu Grafiği), bir veri setindeki dağılımı ve merkezi eğilimi görselleştiren bir grafik türüdür. Veri setindeki minimum, maksimum, medyan, çeyrekler ve aykırı değerleri hızlı bir şekilde analiz etmek için kullanılır. İstatistiksel özetleme amacıyla oldukça yaygın bir şekilde tercih edilir.

aes (): Grafik üzerindeki estetik bileşenleri tanımlar. Bu bileşenler genellikle verilerin hangi ekseninde yer alacağıdır.

aes(x = factor, y = values) şeklinde kullanılır. Burada x eksenine kategorik, y eksenine ise sayısal değişken yerleştirilir.

geom_boxplot (): Verilerin dağılımını görselleştiren boxplot grafik türünü çizer.

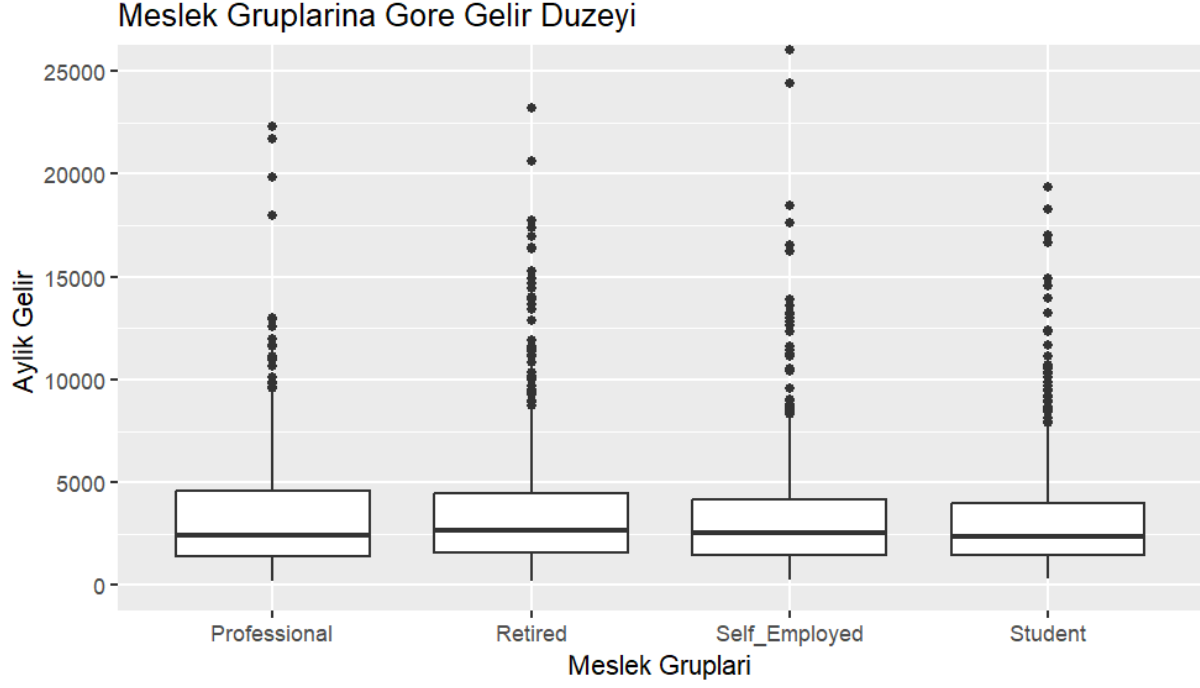
geom_boxplot() yazılarak boxplot tipi grafik oluşturulmuş olur.

labs (): Grafik başlıkları ve eksen etiketlerini tanımlamak için kullanılır.

labs(title = "Başlık", x = "X Eksen Başlığı", y = "Y Eksen Başlığı") şeklinde kullanılır. Burada title başlık, x ve y ise eksen etiketlerini tanımlar.

➤ Meslek gruplarına göre gelir düzeyinin boxplot grafiklerine bakılır.,

```
library(ggplot2)
> ggplot(data) +
+   aes(x = occupation, y =income/12) +
+   geom_boxplot() +
+   labs(
+     title = "Meslek Gruplarına Gore Gelir Duzeyi",
+     x = "Meslek Gruplari",
+     y = "Aylik Gelir "
+   ) + coord_cartesian(ylim = c(0,25000))
> max(income)
[1] 1079728
```



- Bu grafik, meslek gruplarına göre gelir dağılımında farklılıklar olduğunu açıkça ortaya koyuyor. Medyan gelir değerleri meslek grupları arasında büyük fark göstermese de, özellikle aykırı değerler bu meslek grupları içindeki çeşitliliği işaret ediyor.

2.10 Histograma Yoğunluk Tahmini Ekleme

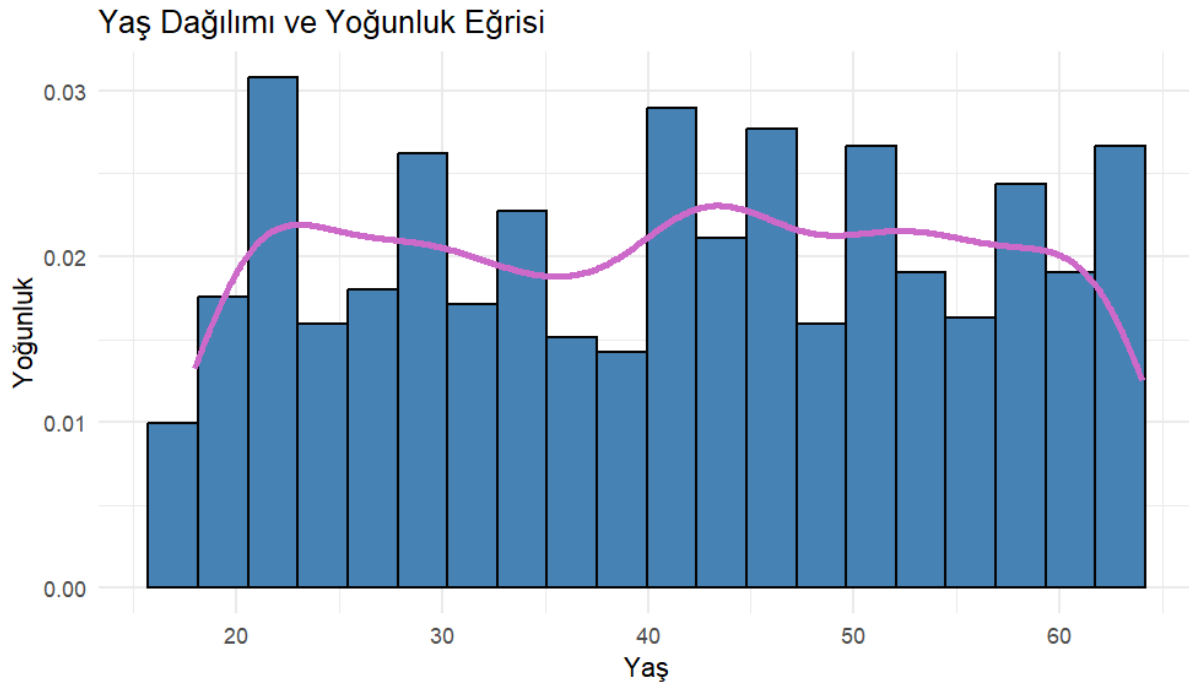
Histogram ve yoğunluk tahmini grafikleri, veri analizi sürecinde değişkenlerin dağılımlarını anlamak, karşılaştırmalar yapmak ve modelleme için uygun istatistiksel yöntemleri seçmek adına kritik öneme sahiptir. Bu grafikler, bir veri setinin temel özelliklerini hızlı ve etkili bir şekilde anlamamızı sağlar.

geom_density() Fonksiyonu:

Fonksiyonunu kullanarak, verinin yoğunluğunu daha düzgün bir şekilde tahmin edebilirsiniz. Histogramın üzerine eklenen bu yoğunluk eğrisi, verinin dağılımını daha net bir şekilde gösterir.

- Yaş değişkeninin dağılımı ve yoğunluk eğrisine bakılır.

```
ggplot() +  
+   aes(x = age) +  
+   geom_histogram(aes(y = ..density..), bins = 20, fill = "steelblue", color  
= "gray2") +  
+   geom_density(color = "orchid3", size = 1.2) +  
+   labs(title = "Yaş Dağılımı ve Yoğunluk Eğrisi",  
+   x = "Yaş",  
+   y = "Yoğunluk") +  
+   theme_minimal()
```



- Grafik, yaş dağılımını ve genel eğilimi göstermektedir. 20-30 yaş aralığında en yüksek yoğunluk gözlenirken, yaş ilerledikçe yoğunluk azalmaktadır. 60 yaş ve üzeri bireyler daha az temsil edilmiştir, genel olarak veri genç ve orta yaş gruplarına odaklanmıştır.

2.11 Normal (Q-Q) Grafiği Oluşturma

QQ grafiği, gözlemlenen gelir verilerinin normal dağılıma ne kadar uygun olduğunu görselleştirir.

- Gelir değişkeninin normal dağılıma uygunluğu test edilir.

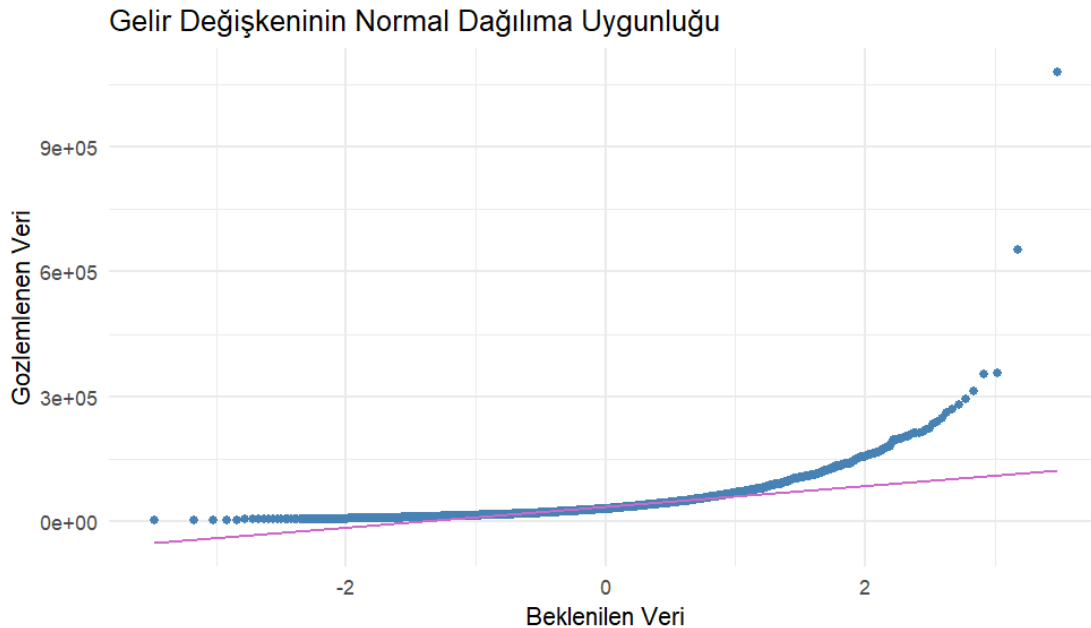
```
> #QQPLOT OLUŞTURMA
> library(dplyr)
> ggplot(data, aes(sample = income)) +
+   stat_qq(color = "steelblue") +
+   stat_qq_line(color = "orchid3") +
+   labs(title = iconv("Gelir Değişkeninin Normal Dağılıma Uygunluğu", to = "UTF-8"),
+   x = ("Beklenilen Veri"),
+   y = ("Gözlemlenen Veri"))+
+   theme_minimal()
> shapiro.test(income)
```

Shapiro-Wilk normality test

data: income
W = 0.57443, p-value < 2.2e-16

Shapiro-Wilk normality testinde $W=0.57443$ ve $p<2.2e-16$ bulunmuştur.

p-değerinin 0.05'ten küçük olması, gelirin normal dağılıma uygun olmadığını istatistiksel olarak doğrular.



- Veriler genellikle normal dağılım çizgisine yakın olmalı, ancak bu grafikte uç oktalar (aşırı büyük gelirler) çizgiden ciddi şekilde sapıyor.
- Bu durum, gelir değişkeninin normal dağılımdan uzak olduğunu gösterir.

2.12 Değişkenleri Farklı Renklerle Görselleştirme

case when() Fonksiyonu : Bir veri kümesinde koşullu bir şekilde yeni değerler atamak için kullanılır.

mutate() Fonksiyonu: Bir veri çerçevesine (data frame) yeni bir sütun eklemek veya mevcut bir sütunu değiştirmek için kullanılır.

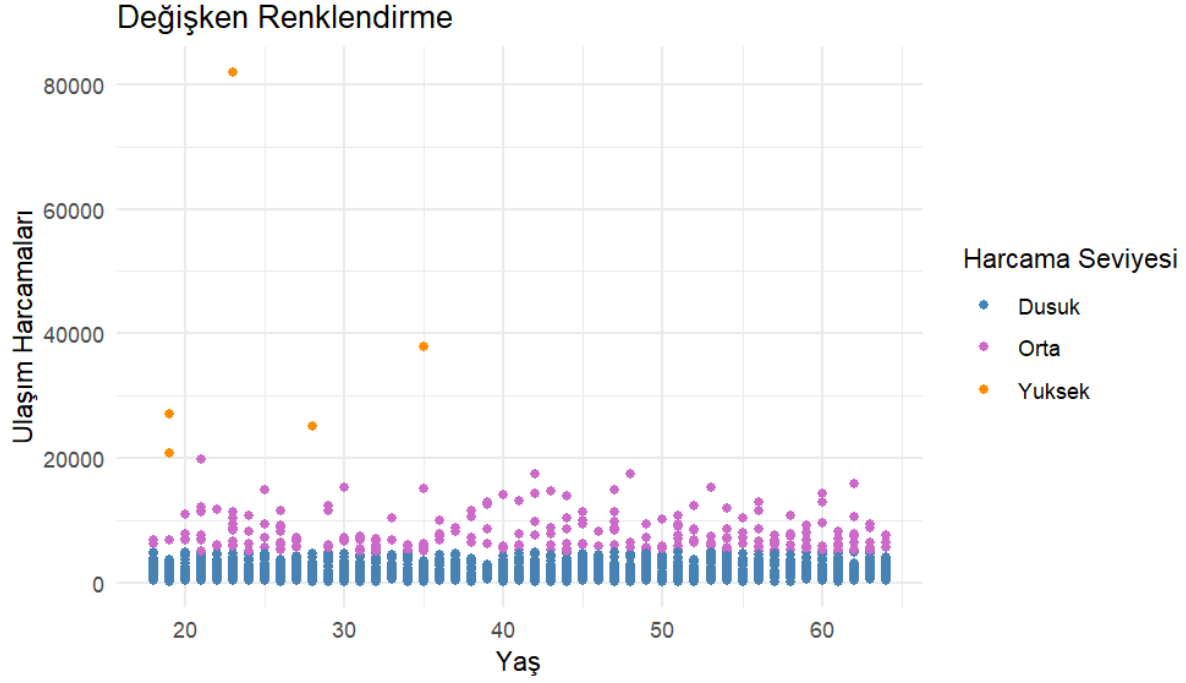
Bir veriyi görselleştirirken, her kategori (örneğin, bir faktör veya grup) için farklı şekiller kullanmak istersiniz. Bu, grafiği daha bilgilendirici ve görsel olarak daha anlaşılır hale getirebilir.

ggplot2'de, `geom_point()` fonksiyonunda `shape` estetiğini kullanarak kategorik bir değişkene göre farklı şekiller atanabilir. Her bir şekil, farklı bir sembol ile temsil edilir (örneğin, daireler, kareler, üçgenler).

`case_when` fonksiyonu, R'deki `dplyr` paketine ait bir fonksiyondur ve veri çerçevelerinde şartlı olarak yeni sütunlar oluşturmak veya mevcut değerleri güncellemek için kullanılır.

- Ulaşım harcamaları değişkeninin yaş değişkenine göre dağılımına bakılır.

```
> library(dplyr)
> data <- data %>%
+   mutate(transport_category = case_when(
+     transport <= 5000 ~ "Dusuk",
+     transport > 5000 & transport <= 20000 ~ "Orta",
+     transport > 20000 ~ "Yuksek"
+   ))
> library(ggplot2)
> ggplot(data) +
+   aes(x = age, y = transport, color = transport_category) +
+   geom_point() +
+   scale_color_manual(values = c("steelblue", "orchid3", "darkorange")) +
+   labs(
+     title = "Değişken Renklendirme",
+     x = "Yaş",
+     y = "Ulaşım Harcamaları",
+     color = "Harcama Seviyesi"
+   ) +
+   theme_minimal()
```



- Genel olarak, ulaşım harcamalarının büyük bir kısmı düşük seviyede yoğunlaşmaktadır. Orta seviyedeki harcamalar, düşük seviyelere kıyasla daha az görülmekte ancak geniş bir yaş aralığında dağılmaktadır. Yüksek harcamalar ise oldukça seyrek olup genellikle daha genç yaş gruplarında gözlenmektedir. Bu durum, ulaşım harcamalarının yaşa göre belirgin bir artış ya da azalma göstermediğini, ancak çoğunlukla düşük seviyede yoğunlaştığını ortaya koymaktadır.

2.13 Basit Doğrusal Regresyon Modeli

lm() Fonksiyonu: Lineer regresyon modelleri oluşturmak için kullanılır. Bu fonksiyon, bir bağımlı değişken ile bir veya daha fazla bağımsız değişken arasındaki doğrusal ilişkiyi modellemeye yarar.

Doğrusal regresyon modeli, iki değişken arasında doğrusal bir ilişki olduğunu varsayar. Bu örnekte x bağımsız ve y bağımlı değişken olarak kabul edilmiştir.

- Eğlence harcamalarının gelir değişkeni tarafından nasıl açıklandığını ifade eder.

```

> model <- lm(entertainment ~ income)
> model

Call:
lm(formula = entertainment ~ income)

Coefficients:
(Intercept)      income
-52.94148      0.03595

```

\hat{y} : entertainment x: income

$$\hat{y} = -52.94148 + 0.03595 \cdot x$$

- Gelir ve eğlence harcamaları arasındaki ilişki pozitif. Gelir arttıkça eğlence harcamalarının da artması beklenir.
- Intercept (Kesme noktası), negatif bir değer olduğu için bu aralık dışında modelin geçerli olmadığı düşünülebilir.

2.14 Çoklu Doğrusal Regresyon

➤ Bu model, gelirin yaş, şehir ve meslek gibi faktörlerle nasıl ilişkilendiğini gösterir.

```

mul_model <- lm(income ~ age + city + occupation)
> mul_model

Call:
lm(formula = income ~ age + city + occupation)

Coefficients:
(Intercept)      age
45094.83      -15.33
cityTier_2      cityTier_3
964.36      48.64
occupationRetired occupationSelf_Employed
-345.48      -2876.81
occupationStudent
-6090.00

```

\hat{y} : Bağımlı değişken (income)

b_0 : Kesme noktası (45094.8345094.8345094.83)

b_1, b_2, \dots, b_k : Katsayılar

x_1, x_2, \dots, x_k : Bağımsız değişkenler (age, city, occupation)

$$\hat{y} = 45094.83 - 15.33 \cdot x_1 + 964.36 \cdot x_2 + 48.64 \cdot x_3 - 345.48 \cdot x_4 - 2876.81 \cdot x_5 - 6090.00 \cdot x_6$$

- Tüm bağımsız değişkenlerin etkisi sıfırken, income için başlangıç değeri 45094.83'tür.
- Yaşın her bir birim artışı, income'da -15.33 birim azalmaya neden olur.
- Eğer şehir Tier_2 ise, gelir $+ 964.36$ artar.
- Eğer şehir Tier_3 ise, gelir $+ 48.64$ artar.
- Eğer kişi Retired (Emekli) ise, gelir $- 345.48$ düşer.
- Eğer kişi Self_Employed (Serbest Meslek) ise, gelir $- 2876.81$ düşer.
- Eğer kişi Student (Öğrenci) ise, gelir $- 6090.00$ düşer.

2.15 Summary

summary() Fonksiyonu:

R'deki summary() fonksiyonu, bir vektör, veri çerçevesi, regresyon modeli veya ANOVA modeli ndeki değerleri hızlıca özetlemek için kullanılabilir.

- **Veri Setleri:** Değişkenlerin temel istatistiklerini (minimum, maksimum, medyan, ortalama vb.) sunar.
- **Regresyon Modelleri:** Modelin katsayıları, standart hataları, p-değerleri ve t-istatistiklerini listeler.
- **Nesneler:** Matris, liste gibi R nesnelerinin uygun bir özetini sunar.

```

> #SUMMARY
> summary(model)

Call:
lm(formula = entertainment ~ income)

Residuals:
    Min       1Q   Median       3Q      Max
-4202.0  -187.5    24.4   208.7  6628.5

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.294e+01  1.586e+01  -3.339 0.000857 ***
income       3.595e-02  2.504e-04 143.538 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 525.1 on 1998 degrees of freedom
Multiple R-squared:  0.9116, Adjusted R-squared:  0.9116
F-statistic: 2.06e+04 on 1 and 1998 DF, p-value: < 2.2e-16

```

➤ Eğlence harcamalarının gelir değişkeni tarafından nasıl açıklandığını ifade eder.

\hat{y} : entertainment x: income

$\hat{y} = -52.94148 + 0.03595 \cdot x$

Artıklar (Residuals):

- Artıklar, tahmin edilen değer ile gözlemlenen değer arasındaki farktır.
- Minimum artık: -4202.0
- Maksimum artık: 6628.5
- Median: 24.4 (orta düzeyde bir sapma olduğunu gösteriyor).

Katsayılar (Coefficients):

- **Intercept:**
 - Tahmini değer: -52.94
 - Standart hata: 15.86
 - t-değeri: -3.339 (istatistiksel olarak anlamlı, çünkü p-değeri 0.000857 < 0.001).

- **income:**
 - Tahmini katsayı: 0.03595
 - Bu, gelirdeki her bir birim artışın, eğlence harcamalarında ortalama 0.03595 birim artışa neden olduğunu gösterir.
 - Standart hata: 0.0002504
 - t-değeri: 143.538 (çok yüksek, bu da bu değişkenin güçlü bir etkisi olduğunu gösterir).
 - p-değeri: $< 2.2 \times 10^{-16}$ (gelirin eğlence üzerinde çok anlamlı bir etkisi olduğunu gösteriyor).

Model Performansı:

- **R-squared (R^2): 0.9116**
 - Gelir değişkeni, eğlence harcamalarındaki değişimin %91,16'sını açıklamaktadır.
- **Adjusted R-squared: 0.9116**
 - R^2 değerinin düzeltilmiş hali, bağımsız değişken sayısına göre uyarlanır.
- **F-istatistiği: 2.06e+04**
 - Modelin genel olarak anlamlı olduğunu gösteriyor (p-değeri $< 2.2 \times 10^{-16}$).
- Modelde, gelir ile eğlence harcamaları arasında güçlü ve pozitif bir ilişki vardır.
- Gelir değişkeni, eğlence harcamalarındaki varyansın büyük kısmını (%91,16) açıklıyor.
- Katsayıların tümü istatistiksel olarak anlamlı ($p < 0.001$).
- Ancak, artıkların geniş bir aralıkta yayılması (−4202 ile 6628.5) modelin bazı gözlemleri iyi tahmin edemediğini gösterebilir. Bu durum, modelin incelenmesini veya başka bağımsız değişkenlerin eklenmesini gerektirebilir.

2.16 Dönüştürülmüş Veri Regresyonu

log() Fonksiyonu: Bir sayının veya bir dizi sayının logaritmasını hesaplamak için kullanılır.

Dönüştürülmüş veri regresyonunun temel amacı, bağımlı ve bağımsız değişkenler arasındaki ilişkiyi daha iyi bir forma sokmak, doğrusal regresyonun varsayımlarını karşılamak ve daha güvenilir model sonuçları elde etmektir.

- Gelir değişkeninin logaritmik dönüşümü ile dışarıda yemek değişkeni arasındaki basit doğrusal regresyon modelini gösterir.

$$\log(\hat{y}) = -52.94148 + 0.03595 \cdot x$$

\hat{y} : income x: eating_out

```
> transformed_model <- lm(log(income) ~ eating_out)
> transformed_model

Call:
lm(formula = log(income) ~ eating_out)

Coefficients:
(Intercept)      eating_out 
  9.7422610      0.0003961 

> summary(transformed_model)

Call:
lm(formula = log(income) ~ eating_out)

Residuals:
    Min       1Q   Median       3Q      Max 
-4.9457 -0.2471  0.0730  0.3098  0.9636 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.742e+00  1.511e-02  644.79  <2e-16 ***
eating_out    3.961e-04  6.987e-06   56.69  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4934 on 1998 degrees of freedom
Multiple R-squared:  0.6166, Adjusted R-squared:  0.6165 
F-statistic: 3214 on 1 and 1998 DF, p-value: < 2.2e-16
```

Artıklar (Residuals):

- Artıkların aralığı: -4.9457 ile 0.9636
- Artıkların medyanı: 0.073 (medyan sifıra yakın, model dengeli artıklar üretmiş).

Katsayılar (Coefficients):

- **Intercept:**
 - Katsayı: 9.742
 - Standart hata: 0.01511
 - t-değeri: 644.79 (çok yüksek, istatistiksel olarak anlamlı).
 - p-değeri: $<2e-16$ (çok anlamlı).
- **eating_out:**
 - Katsayı: 0.0003961
 - Bu, dışarıda yeme harcamalarındaki her bir birim artışın logaritmik gelirden ortalama 0.0003961 birim artışa neden olduğunu gösterir.
 - t-değeri: 56.69 (yüksek anlamlılık).
 - p-değeri: $<2e-16$ (istatistiksel olarak çok anlamlı).

Model Performansı:

- **Residual Standard Error (RSE):** 0.4934
 - Artıkların standart sapmasıdır; modelin tahmin ettiği değerler ile gerçek değerler arasındaki sapmayı ölçer.
- **Multiple R-squared (R^2):** 0.6166
 - **eating_out** değişkeni, logaritmik gelir değişiminin %61.66'sını açıklıyor.
- **Adjusted R-squared:** 0.6165
 - R^2 'nin düzeltilmiş hali, modelin bağımsız değişken sayısına göre uyarlanmış versiyonudur.

- **F-statistik:** 3214, p-değeri $<2.2e-16$
 - Modelin genel olarak anlamlı olduğunu ve **eating_out** değişkeninin logaritmik gelir üzerinde güçlü bir etkisi olduğunu gösterir.
- **eating_out** harcamaları ile logaritmik gelir arasında pozitif ve anlamlı bir ilişki var.
- **eating_out**, logaritmik gelir değişkenindeki varyansın %61.66'sını açıklayabiliyor, bu oldukça güçlü bir açıklama oranı.
- Artıklar, modelin bazı gözlemleri iyi tahmin edemediğini gösterebilir (örneğin, minimum artık oldukça düşük).
- Logaritmik dönüşüm, modelin performansını artırmış olabilir ve bağımlı değişkendeki doğrusal olmayan ilişkileri yakalamış olabilir.

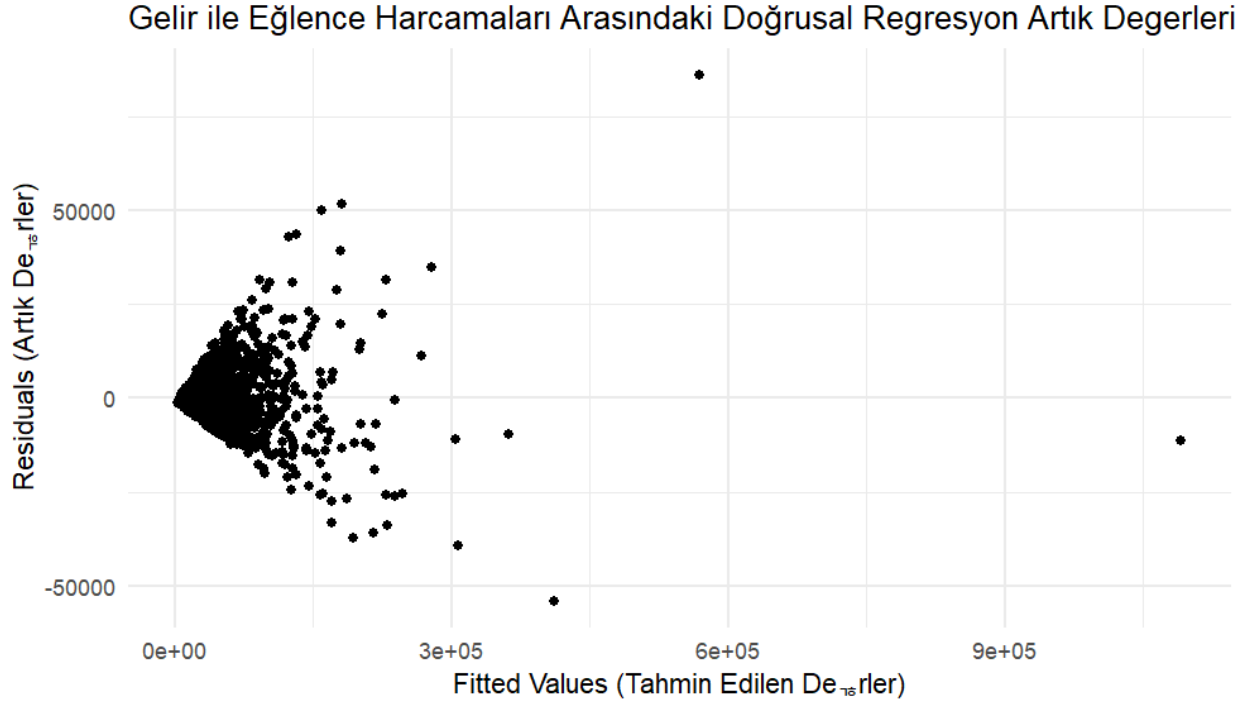
Model, bağımsız değişkenin logaritmik gelir üzerindeki etkisini başarılı bir şekilde göstermektedir.

2.17 Regresyon Modelinin Artık Değerlerini Görselleştirme

augment() Fonksiyonu: Bir istatistiksel modelin tahmin sonuçlarını, artıklarını (residuals), tahmin edilen değerlerini (fitted values) ve modelle ilişkili diğer bilgileri içeren bir veri çerçevesi döndürür.

broom paketi, R'da modelleme sonuçlarını özetlemek ve daha anlaşılır hale getirmek için kullanılan bir pakettir. Bu paket, model çıktısını düzenli formatta veri çerçevesi olarak sağlar. Bu, modelleme sonuçlarını analiz etmek, görselleştirmek veya başka işlemler için kolayca kullanılmasına olanak tanır.

```
> res <- lm(income ~ eating_out + transport )
> library(broom)
> augmented_res <- augment(res)
> library(ggplot2)
> ggplot(augmented_res, aes(x = .fitted, y = .resid)) +
+   geom_point() +
+   labs(title = "Gelir ile Eğlence Harcamaları Arasındaki Doğrusal Regresyon Artık Değerleri",
+         x = "Fitted Values (Tahmin Edilen Değerler)",
+         y = "Residuals (Artık Değerler)") +
+   theme_minimal()
```



➤ Gelir ve eğlence harcamaları arasındaki doğrusal regresyonun artık değerleri görselleştirilir.

- Artık değerler rastgele bir desen göstermeli ve sabit varyansa sahip olmalıdır. Bu grafikte, varyansın sabit olmaması ve uç değerlerin varlığı, model varsayımlarında problemlerin olduğunu işaret ediyor.
- Değişken varyansı azaltmak için log dönüşümü gibi bir dönüşüm uygulanabilir. Ayrıca uç değerlerin modelden çıkarılması veya ayrı bir incelemeye tabi tutulması gerekebilir.

2.18 Etkili Gözlemleri Tanımlama

influence.measures() Fonksiyonu: Bir regresyon modeli üzerindeki veri noktalarının etkisini analiz etmek için kullanılır.

- **dfb (Difference in Beta):** Modelin katsayılarındaki değişiklikleri ifade eder. Her değişken için ayrı ayrı hesaplanır. Bir gözlemin, modeldeki katsayıları ne ölçüde etkilediğini gösterir.
- **DFFITS (Difference in Fits):** Bir gözlemin modele etkisini değerlendirir. Yüksek mutlak değerler, gözlemin tahmin değerleri üzerinde önemli bir etkisi olduğunu gösterir. (dffit)
- **Covariance Ratio:** Modeldeki gözlemin varyans-kovaryans matrisine etkisini ölçer. (cov.r)
- **Cook's Distance:** Bir gözlemin, tahmin edilen modelin genel sonuçlarını nasıl etkilediğini ölçen bir istatistik. (cook.d)
- **Hat (Leverage):** Bir gözlemin, modelin tahmin edilebilirliğine olan etkisini gösterir. 1'e yakın değerler, yüksek kaldıraç anlamına gelir. (hat)
- Yıldız (*) sembolü, potansiyel olarak etkili gözlemleri işaretler.

```
> model1 <- lm(income ~ rent + transport )
> influential_obs <- influence.measures(model1)
> summary(influential_obs)
Potentially influential observations of
lm(formula = income ~ rent + transport) :
```

	dfb.1_	dfb.rent	dfb.trns	dffit	cov.r	cook.d	hat
15	0.01	0.06	-0.08	-0.08	1.02_*	0.00	0.02_*
20	0.00	0.01	-0.01	0.01	1.00_*	0.00	0.00
30	0.02	0.04	-0.06	-0.08	1.00_*	0.00	0.01_*
55	-0.01	0.03	-0.03	0.03	1.01_*	0.00	0.01_*
58	-0.01	0.07	-0.04	0.12_*	0.99_*	0.01	0.00
61	0.00	-0.02	0.01	-0.02	1.00_*	0.00	0.00
65	0.00	-0.02	0.02	-0.02	1.01_*	0.00	0.01_*
111	-0.07	0.17	-0.10	0.27_*	0.98_*	0.02	0.00
125	-0.05	0.06	-0.02	0.16_*	0.99_*	0.01	0.00
127	0.01	0.06	-0.04	0.09	0.99_*	0.00	0.00
129	0.00	-0.09	0.11	0.13_*	1.00	0.01	0.00
141	0.00	-0.01	0.01	-0.01	1.00_*	0.00	0.00

```
.....
.....
.....
```

2.19 One-way ANOVA

oneway.test () Fonksiyonu: Gruplar arasındaki ortalama farklılıklarını test etmek için kullanılan etkili bir fonksiyondur. Welch'in ANOVA'sını desteklediği için eşit varyans varsayımına gerek duymayan durumlarda idealdir.

Tek yönlü ANOVA (One-Way ANOVA) testi, gruplar arasında varyansların eşit olmadığını varsayarak gerçekleştirilmiştir.

- Şehir düzeyine göre ulaşım masrafları arasında anlamlı bir fark olup olmadığını incelemek için tek yönlü ANOVA (Analysis of Variance) testi yapılır.

```
oneway <- oneway.test(transport ~ city)
> oneway

One-way analysis of means (not assuming equal variances)

data: transport and city
F = 0.26273, num df = 2.0, denom df = 1160.1, p-value = 0.769
```

- ANOVA testindeki p-değeri, şehir düzeyine göre ulaşım masrafları arasında anlamlı bir fark olup olmadığını değerlendirir. p-değeri 0.769, %5 anlamlılık düzeyinden ($\alpha = 0.05$) oldukça yüksektir. Bu, şehir düzeyine göre ulaşım masrafları arasında istatistiksel olarak anlamlı bir fark olmadığını gösterir.
- F-değeri, gruplar arası varyansın grup içi varyansa oranını temsil eder. Düşük bir F-değeri, gruplar arasında belirgin bir fark olmadığını gösterir. Bu testteki F-değeri (0.26273), gruplar arasındaki farkın istatistiksel olarak anlamlı olmadığını destekler.
- p-değerinin yüksek olması ve F-değerinin düşük olması nedeniyle, şehir düzeyine göre ulaşım masrafları arasında anlamlı bir fark olmadığı sonucuna ulaşılmıştır.

3.SONUÇ

Bu rapor sonucunda bireylerin gelir düzeyleri, harcama alışkanlıkları ve demografik faktörler arasındaki ilişkiler detaylı olarak incelenmiştir.

Gelir ve Harcama İlişkisi: Gelir düzeyi, bireylerin farklı harcama kategorilerindeki tutumlarını etkilemektedir. Örneğin, gelir arttıkça eğlence ve dışarıda yemek gibi isteğe bağlı harcamalarda da artış gözlemlenmiştir.

Demografik Faktörlerin Etkisi: Yaş ve meslek gibi demografik faktörler, harcama alışkanlıklarında farklılık yaratmaktadır. Genç bireyler (18-25 yaş grubu) gelirlerinin yüksek olmasına rağmen, nispeten daha fazla isteğe bağlı harcama yapmaktadır. Emekliler ise ulaşım daha fazla harcama ayırmaktadır.

Korelasyonlar: Market ve dışarıda yemek harcamaları arasında pozitif bir ilişki bulunmuştur. Bu, bireylerin tüketim alışkanlıklarının genelde birbiriyle paralel olduğunu göstermektedir.

Regresyon Modelleri: Çoklu regresyon modeli, yaş, şehir düzeyi ve meslek gibi faktörlerin gelir üzerindeki etkisini anlamada etkili bir araç olmuştur. Ancak modeldeki artık değerlerin geniş bir aralıkta yayılması, modelin bazı durumlarda yetersiz kaldığını göstermiştir.

4. KAYNAKÇA

1. Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
Erişim Adresi: <https://ggplot2-book.org/introduction#:~:text=ggplot2%20is%20an%20R%20package,This%20makes%20ggplot2%20powerful>
2. RC2E. (n.d.).
Erişim Adresi: <https://rc2e.com/>
3. R Markdown. (n.d.). RStudio.
Erişim Adresi: <https://rmarkdown.rstudio.com/lesson-1.html>
4. İstanbul Teknik Üniversitesi Matematiksel ve Teknolojik Araştırma Laboratuvarı. (n.d.).
Erişim Adresi: <https://www.itumtal.com/belgeler/veri/>
5. Statology. (n.d.).
Erişim Adresi: <https://www.statology.org/>

	Kodların Yazımı	Analiz	Yorumlama
Hilal KAYA	X	X	X
Beyza Senanur AKTAŞ	X	X	X