

```
library(corr)  
library(tidyverse)  
library(nycflights13)
```

```
# View(planes)
```

Soru1. Havada kalış süresi (air_time) en fazla olan uçak için üretim yılı (year), üretici (manufacturer) ve model (model) bilgisi nedir? (Kodunuzda hata alıyorsanız year değişkenini verinizde kontrol ediniz!)

```
data1<- left_join(flights,planes,by="tailnum")
```

```
data1 %>% filter(air_time==max(air_time,na.rm = T)) %>%  
  select(year,y, manufacturer,model)
```

```
→ # A tibble: 1 × 3
```

```
  year.y manufacturer model
```

```
  <int> <chr>      <chr>
```

```
1  2002 BOEING      767-424ER
```

Soru2. Havada kalış süresi (air_time), sıcaklık (temp), rüzgar hızı (wind_speed) ve nem (humid) arasında ne tür bir ilişki vardır? İnceleyip yorumlayınız.

```
data2<- left_join(weather,flights,by=c("origin","year","month","day","hour"))  
data2 %>% select(air_time,temp,wind_speed,humid) %>%  
  correlate()
```

```
→ # A tibble: 4 × 5
```

```
  term  air_time  temp wind_speed  humid
```

```
  <chr>   <dbl>  <dbl>   <dbl>  <dbl>
```

```
1 air_t... NA    -0.0367  0.0263  0.0405
```

```
2 temp  -0.0367 NA     -0.140  0.0374
```

```
3 wind_... 0.0263 -0.140   NA     -0.187
```

```
4 humid  0.0405 0.0374 -0.187  NA
```

Soru3. Her bir hava yolu şirketi (name) için ortalama ve ortanca gecikme süresilerini (dep_delay) inceleyip yorumlayınız.

```
data3<- right_join(airlines,flights,by="carrier")
```

```
data3 %>% filter(dep_delay>0) %>% na.omit() %>% group_by(name) %>%
summarise(mean=mean(dep_delay),median=median(dep_delay))
```

```
# A tibble: 16 × 3
```

	name	mean	median
	<chr>	<dbl>	<dbl>
1	AirTran Airways Corporation	40.6	16
2	Alaska Airlines Inc.	31.5	12
3	American Airlines Inc.	37.2	16
4	Delta Air Lines Inc.	37.3	16
5	Endeavor Air Inc.	48.5	26
6	Envoy Air	44.7	27
7	ExpressJet Airlines Inc.	50.2	31
8	Frontier Airlines Inc.	45.2	18
9	Hawaiian Airlines Inc.	44.8	5
10	JetBlue Airways	39.7	20
11	Mesa Airlines Inc.	52.9	29.5
12	SkyWest Airlines Inc.	58	40
13	Southwest Airlines Co.	34.8	15
14	US Airways Inc.	32.9	16
15	United Air Lines Inc.	29.8	12
16	Virgin America	34.2	10

Soru4. flights veri setinde yer alan gecikme süresi (dep_delay) değişkeni için 2013 yılı ortalama gecikme süresi kaçtır? (erken varışları dikkate almayınız.)

```
flights %>% filter(dep_delay>0) %>% summarise(mean=mean(dep_delay))
```

```
→# A tibble: 1 × 1
```

	mean
	<dbl>
1	39.4

```
-----
library(Lahman)
library(tidyverse)
```

Soru1. AwardsPlayers veri setini incelediğinizde en çok ödül alan oyuncu kimdir? (People veri seti nameFirst ve nameLast değişkenleri ile yanıtlayınız.)

```
t<-table(AwardsPlayers$playerID) %>% as.data.frame()
names(t)<-c("playerID","num.of.aw")
data1<-People %>% left_join(t,by="playerID") # hepsine na.omit()
uygulamak?
data1 %>% filter(num.of.aw==max(num.of.aw,na.rm=T)) %>%
  select(nameFirst,nameLast)
```

```
→ nameFirst nameLast
1 Barry Bonds
```

Soru2. People veri setinde yer alan weight ve height değişkenlerini kullanarak body mass index (BMI) hesaplayınız. BMI formülü ve değerlendirme tablosu aşağıda yer almaktadır.

```
People <- data.frame(
  weight = c(150, 180, 200, 120, 160), # pound cinsinden ağırlık
  height = c(65, 70, 72, 60, 68) # inç cinsinden boy
)
```

BMI hesaplama

```
People$BMI <- (People$weight / (People$height^2)) * 703
```

Sınıflandırma

```
People$BMI_class <- cut(People$BMI,
  breaks = c(-Inf, 18.5, 24.9, 29.9, 34.9, 39.9, Inf),
  labels = c("Under weight", "Normal weight", "Over weight",
    "Obesity class 1", "Obesity class 2", "Obesity class 3"))
```

Sonuçları yazdırma

```
print(People)
```

```
→ weight height BMI BMI_class
1 150 65 24.95858 Over weight
2 180 70 25.82449 Over weight
3 200 72 27.12191 Over weight
4 120 60 23.43333 Normal weight
5 160 68 24.32526 Normal weight
```

Oyunculardan kaç tanesi normal weight kategorisine girmektedir?

```
People %>% mutate(BMI=weight/(height^2)*703) %>% filter(BMI>=25 &
  BMI<29.9) %>% nrow()
```

→ [1] 2

Soru3. Salaries ve AwardsPlayers veri setlerini inceleyiniz, en fazla ödül alan oyuncunun en yüksek maaşa sahip olduğu söylenebilir mi?
#soru 1 de odul sayilari hesaplandi (t)

```
data2<-Salaries %>% left_join(t,by="playerID")
which.max(data2$salary)==which.max(data2$num.of.aw)
slice(data2,which.max(data2$salary)) #max salary
slice(data2,which.max(data2$num.of.aw)) #max num of awards
```

→ [1] FALSE

```
yearID teamID lgID playerID salary
1 2009 NYA AL rodrial01 33000000
num.of.aw
1 31
yearID teamID lgID playerID salary
1 1986 PIT NL bondsba01 60000
num.of.aw
1 47
```

Soru4. En çok alınan ödül hangisidir?

```
table(AwardsPlayers$awardID) %>% as.data.frame() %>%
arrange(desc(Freq))
```

```
→
      Var1 Freq
1      TSN All-Star 1525
2  Baseball Magazine All-Star 1520
3      Gold Glove 1204
4    Silver Slugger 792
5  Most Valuable Player 208
6  Rookie of the Year 154
7  TSN Pitcher of the Year 151
8      Cy Young Award 126
9  Reliever of the Year Award 94
10 TSN Player of the Year 92
11 TSN Major League Player of the Year 89
12 TSN Fireman of the Year 88
13    Babe Ruth Award 78
```

14	World Series MVP	71
15	Lou Gehrig Memorial Award	69
16	All-Star Game MVP	62
17	Hutch Award	55
18	Roberto Clemente Award	54
19	Hank Aaron Award	50
20	NLCS MVP	49
21	ALCS MVP	43
22	Pitching Triple Crown	39
23	Comeback Player of the Year	36
24	TSN Reliever of the Year	36
25	TSN Guide MVP	33
26	Platinum Glove	26
27	Branch Rickey Award	23
28	Triple Crown	17
29	Outstanding DH Award	8
30	Silver Slugger	5

```
library(ggplot2)
library(tidyverse)
```

Soru1. diamonds veri setinde yer alan x, y ve z değişkenleri kullanılarak yeni bir t değişkeni oluşturuluyor. t değişkeni $t = x^2 - \sqrt{y} + z^{(-2)}$ şeklinde tanımlanacak olursa, en düşük t değerine karşılık gelen depth değeri nedir?

```
diamonds %>% mutate(t=x^2-sqrt(y)+(1/z)) %>% filter(t==min(t)) %>%
select(depth)
```

→ # A tibble: 1 × 1

```
depth
```

```
<dbl>
```

```
1 62.6
```

A tibble: 1 × 12

```
carat cut    color clarity depth table
```

```
<dbl> <ord> <ord> <ord> <dbl> <dbl>
```

```
1 0.23 Ideal E    SI2    61.5 55
```

```
# i 6 more variables: price <int>, x <dbl>,
```

```
# y <dbl>, z <dbl>, discount <dbl>,
```

```
# new.price <dbl>
```

Soru2. Pandemi nedeniyle azalan satışlarını hareketlendirmek isteyen bir mağaza sahibi ürünlerine cut değişkenine bağlı olarak indirim uyguluyor. İndirim oranları aşağıdaki gibidir. İndirimli fiyatlar üzerinden alışveriş yapan bir müşteri 'E color' ve 'SI1 clarity' ürün için en az kaç \$ ödeyecektir?

```
diamonds %>% mutate(discount = case_when(
  cut=="Fair" ~ price*0.01,
  cut=="Good" ~ price*0.02,
  cut=="Very Good" ~ price*0.025,
  cut=="Premium" ~ price*0.03,
  cut=="Ideal" ~ price*0.03,
),new.price=price-discount) %>% filter(color=="E" & clarity=="SI2" &
new.price==min(new.price))
```

→# A tibble: 1 × 12

```
carat cut  color clarity depth table
<dbl> <ord> <ord> <ord> <dbl> <dbl>
1 0.23 Ideal E    SI2    61.5  55
# i 6 more variables: price <int>, x <dbl>,
# y <dbl>, z <dbl>, discount <dbl>,
# new.price <dbl>
```

Alternatif

```
diamonds %>% mutate(discount= case_when(
  cut=="Fair" ~ price*0.1,
  cut=="Good" ~ price*0.12,
  cut=="Very Good" ~ price*0.15,
  cut=="Premium" ~ price*0.18,
  cut=="Ideal" ~ price*0.18,
),new.price=price-discount) %>% filter(color=="E"& clarity=="SI1") %>%
arrange(new.price)
```

→ # A tibble: 2,426 × 12

```
carat cut  color clarity depth table
<dbl> <ord> <ord> <ord> <dbl> <dbl>
1 0.21 Premium E    SI1    59.8  61
2 0.26 Very Good E    SI1    62    54
3 0.28 Good    E    SI1    62.6  60
4 0.27 Very Good E    SI1    61.2  57
5 0.31 Premium E    SI1    62.7  58
```

```

6 0.24 Ideal E SI1 62.5 56
7 0.3 Ideal E SI1 62.1 58
8 0.3 Ideal E SI1 61.1 57
9 0.3 Ideal E SI1 62.4 57
10 0.3 Ideal E SI1 61.6 54
# i 2,416 more rows
# i 6 more variables: price <int>, x <dbl>,
# y <dbl>, z <dbl>, discount <dbl>,
# new.price <dbl>
# i Use `print(n = ...)` to see more rows

```

**# Soru3. diamonds verisi clarity değişkeni kaç farklı level içermektedir?
(Kaç farklı clarity tipi vardır?)**

```

cl<-diamonds$clarity %>% as.factor()
cl %>% levels()
cl %>% nlevels()

```

```

→ [1] "I1" "SI2" "SI1" "VS2" "VS1"
[6] "VVS2" "VVS1" "IF"
[1] 8

```

Soru4. Mağazanın 'Ideal' cut tipine sahip ürünlerden elde ettiği ortalama kazanç nedir?

```

diamonds %>% group_by(cut) %>% summarise(mean.pr=mean(price))

```

```

→ # A tibble: 5 × 2
  cut      mean.pr
<ord>    <dbl>
1 Fair    4359.
2 Good    3929.
3 Very Good 3982.
4 Premium 4584.
5 Ideal   3458.

```

Alternatif

```

diamonds %>% filter(cut=="Ideal") %>% summarise(mean(price))

```

```

→ # A tibble: 1 × 1
  `mean(price)`
    <dbl>

```

1 3458.

Soru5. Mağazanın cut tipine göre elde ettiği price ortalamaları incelendiğinde en az kazanç sağlayan cut tipi hangisidir?

dplyr paketini yükleyin

library(dplyr)

Örnek veri seti (Mağaza verisi)

'cut' ve 'price' sütunları olduğunu varsayalım

store_data <- data.frame(

cut = c("Fair", "Good", "Very Good", "Fair", "Good", "Excellent", "Fair",
"Excellent", "Very Good", "Good"),

price = c(300, 500, 700, 250, 600, 1000, 320, 1100, 750, 550)
)

Her bir 'cut' tipi için fiyat ortalamalarını hesaplayın

average_price_by_cut <- store_data %>%

group_by(cut) %>%

summarise(average_price = mean(price))

En düşük ortalama fiyata sahip 'cut' tipini bulma

lowest_avg_price_cut <- average_price_by_cut %>%

filter(average_price == min(average_price))

Sonuçları yazdırma

print(lowest_avg_price_cut)

→ # A tibble: 1 × 2

cut average_price

<chr> <dbl>

1 Fair 290