Denver Broncos - Assessment

Karl Mbouombouo

2023-07-28

Explore & Clean Dataset

```
procedures <- read_csv("procedures.csv")
```

```
## Rows: 41 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): PetID, ProcedureType
## dbl (2): Date, ProcedureSubCode
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
procedure_details <- read_csv("procedure_details.csv")
```

```
## Rows: 41 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): ProcedureType, Description
## dbl (2): ProcedureSubCode, Price
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
pets <- read_csv("pets.csv")
```

```
## Rows: 100 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): PetID, Name, Kind, Gender
## dbl (2): Age, OwnerID
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
owners <- read_csv("owners.csv")
```

```
## Rows: 89 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (6): Name, Surname, StreetAddress, City, State, StateFull
## dbl (2): OwnerID, ZipCode
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
#Missing Values (NA)
which(is.na(procedures))
which(is.na(procedure_details))
which(is.na(pets))
which(is.na(owners))
```

What is the name of the oldest dog in Southfield?

```r
southfield_pets <- inner_join(owners, pets, by = "OwnerID")

# Filter the data to include only pets in Southfield
southfield_pets <- southfield_pets %>%
  filter(City == "Southfield" & Kind == "Dog")

# Find the oldest dog in Southfield
oldest_dog <- southfield_pets %>%
  filter(Age == max(Age))

# Extract the name of the oldest dog
oldest_dog_name <- oldest_dog$Name.y

# Print the name of the oldest dog
print(oldest_dog_name)
```

```
## [1] "Crockett"
```

The name of the oldest dog is Crockett

What is the average (mean) number of pets per city?

```r
owners_pets <- merge(owners, pets, by = "OwnerID", all.x=TRUE, all.y=FALSE)

# Count the number of pets in each city
pets_per_city <- owners_pets %>%
  group_by(City) %>%
  summarise(total_pets = n())

# Calculate the average number of pets per city
average_pets_per_city <- mean(pets_per_city$total_pets)

# Print the result
print(average_pets_per_city)
```

```
## [1] 2.5
```

2.5

Which owner spent the most on procedures for their pet(s)?

```r
# Combine the data
df_2 <- inner_join(owners, pets, by = "OwnerID") %>%
  inner_join(procedures, by = "PetID") %>%
```

```r
  inner_join(procedure_details, by = "ProcedureSubCode")

# Calculate the total amount spent on procedures for each owner
total_spent_per_owner <- df_2 %>%
  group_by(OwnerID, Name.x, Surname) %>%
  summarise(total_spent = sum(Price))
```

```
## `summarise()` has grouped output by 'OwnerID', 'Name.x'. You can override using
## the `.groups` argument.
```

```r
# Find the owner who spent the most on procedures
owner_with_max_spending <- total_spent_per_owner %>%
  filter(total_spent == max(total_spent))

# Print the result
print(owner_with_max_spending)
```

```
## # A tibble: 33 x 4
## # Groups:   OwnerID, Name.x [33]
##    OwnerID Name.x      Surname   total_spent
##      <dbl> <chr>       <chr>           <dbl>
##  1    1070 Jessica     Velazquez        1302
##  2    1546 Joseph      Blow              665
##  3    1653 Carolyn     Crane             665
##  4    1766 Doris       Ray               236
##  5    1915 Christopher Bowers            665
##  6    2419 Luisa       Cuellar           236
##  7    2722 John        Jordan            665
##  8    2809 Bruce       Dunne            1330
##  9    3086 Ed          Enriquez          665
## 10    3089 Lee         McKenzie         2660
## # i 23 more rows
```

Lee Mckenzie

How many owners spent 20 dollars or more on procedures for their pets?

```r
# Calculate the total cost of procedures for each pet based on 'ProcedureType' and 'ProcedureSubCode'
total_cost_per_pet <- procedure_details %>%
  left_join(procedures, by = c("ProcedureType", "ProcedureSubCode")) %>%
  group_by(PetID) %>%
  summarize(TotalCost = sum(Price, na.rm = TRUE))

# Merge the 'pets' and 'total_cost_per_pet' data on PetID
merge_1 <- merge(pets, total_cost_per_pet, by = "PetID")

# Calculate the total procedure cost per owner
total_cost_per_owner <- merge_1 %>%
  group_by(OwnerID) %>%
  summarize(TotalCost = sum(TotalCost))

# Filter and count the number of owners who spent 20 dollars or more on procedures
```

```r
spent_20_or_more <- total_cost_per_owner %>%
  filter(TotalCost >= 20) %>%
  nrow

print(spent_20_or_more)
```

```
## [1] 12
```

12

How many owners have at least two different kinds of pets?

```r
# Group by 'OwnerID' and count the distinct number of 'Kind' values for each owner
multiple_kinds <- pets %>%
  group_by(OwnerID) %>%
  summarize(NumDifferentKinds = n_distinct(Kind))

# Filter the data to include only the owners with at least two different kinds of pets
owners_with_multiple_kinds <- multiple_kinds %>%
  filter(NumDifferentKinds >= 2) %>%
  nrow()

# Print the result
print(owners_with_multiple_kinds)
```

```
## [1] 4
```

4

How many owners have pets where the first letter of their name (OwnerName) matches the first letter of their pet's name (PetName)?

```r
df_3 <- inner_join(owners, pets, by = "OwnerID")

# Count the number of owners where the first letter of their name matches the first letter of their pet
matching_first_letter <- df_3 %>%
  filter(substr(Name.x, 1, 1) == substr(Name.y, 1, 1)) %>%
  summarise(num_owners = n_distinct(OwnerID))

# Print the result
print(matching_first_letter$num_owners)
```

```
## [1] 8
```

8

What percentage of cities have more male pets than female pets

```r
df_4 <- inner_join(pets, owners, by = "OwnerID")
```

```r
# Calculate the total number of male pets and female pets in each city
pet_count_by_city <- df_4 %>%
  group_by(City) %>%
  summarise(total_male_pets = sum(Gender == "male"),
            total_female_pets = sum(Gender == "female"))

# Count the number of cities where the total number of male pets is greater than the total number of fe
cities_with_more_male_pets <- pet_count_by_city %>%
  filter(total_male_pets > total_female_pets) %>%
  nrow()

# Calculate the percentage of cities with more male pets than female pets
total_cities <- nrow(pet_count_by_city)
percentage_more_male_pets <- (cities_with_more_male_pets / total_cities) * 100

# Print the result
print(percentage_more_male_pets)
```

```
## [1] 57.5
```

57.5

What percentage of pets received a vaccination?

```r
df_5 <- inner_join(pets, procedures, by = "PetID")

# Count the number of pets that received a vaccination
pets_with_vaccination <- df_5 %>%
  filter(ProcedureType == "VACCINATIONS") %>%
  nrow()

# Calculate the percentage of pets that received a vaccination
total_pets <- nrow(pets)
percentage_vaccinated <- (pets_with_vaccination / total_pets) * 100

# Print the result
print(percentage_vaccinated)
```

```
## [1] 29
```

30% (Approximation)

Which city's pet sample is made up of exactly 70% dogs? The answer is case sensitive, so please match the value for City exactly.

```r
# Combine data from 'table 3' and 'table 4' based on 'OwnerID'
df_6 <- pets %>%
  left_join(owners, by = "OwnerID")

# Filter rows where 'Kind' is 'Dog', group by 'City', and count the number of dogs in each city
# Group by 'City' and calculate the total number of pets in each city
result <- df_6 %>%
```

```r
  group_by(City) %>%
  summarize(Num_Dogs = sum(Kind == "Dog"),
            Total_Pets = n())

# Calculate the ratio of dogs to total pets
result$Dog_Ratio <- (result$Num_Dogs / result$Total_Pets) * 100

# Display the table
filtered_70<- filter(result, Dog_Ratio == 70)
print(filtered_70)
```

```
## # A tibble: 1 x 4
##   City         Num_Dogs Total_Pets Dog_Ratio
##   <chr>           <int>      <int>     <dbl>
## 1 Grand Rapids        7         10        70
```

Grand Rapids