# Analiza i klasteryzacja zbioru koktajli

Aleksandr Shestakov
nr index 272657
Politechnika Wrocławska
Wydział W4N

27 October 2024

# Spis treści



# 1 Cele projektu

Celem projektu jest przeprowadzenie preprocessingu, augmentacji i EDA podanego zbioru koktajli oraz klasteryzacji tego zbioru.

# 2 Inital data analysis

Startowy zbiór danych zawiera 134 wiersza i 11 kolumn, co odpowiada 134 koktajlam.

## 2.1 Cechy

Startowy zbiór zawiera kolejne cechy:

- id
- name - nazwa koktajlu
- category - kategoria (Ordinary Drink, Cocktail lub Punch / Party Drink)
- glass - szklanka do serwowania koktajla, spis zostanie podany później
- tags
- instructions - instrkucja do przyrządzenia koktajla

- imageUrl
- alcoholic - czy koktajl jest alkoholowy
- createdAt
- updatedAt
- ingredients - lista słowników ze składnikami koktajla

## 2.2 Wyniki IDA

:

- Wszystkie koktajli okazały się alkoholowymi
- 99 koktajli nie mają danych o ich tag'ach
- createdAt i updatedAt - daty który nie różnią się między sobą i jak podejrzę są momentami w których dane zostali wyciągnięty z TheCokctailDB
- Duplikatów koktajli nie okazało się - wszystkie wierszy dotyczą unikatowych koktajli

Kolumny id, imageUrl, alcoholic, createdAt, updatedAt - zostały wyrzucone jako niepotrzebne do analizy. Ostateczny wynik po tym etapie:

```
cocktails.sample(10)
```

| | name | category | glass | tags | instructions | ingredients |
|---|---|---|---|---|---|---|
| 15 | Brandy Alexander | Ordinary Drink | Whiskey Glass | [Nutty, Dairy] | Shake all ingredients (except nutmeg) with ice... | [{'id': 74, 'name': 'Brandy', 'description': '... |
| 12 | Alabama Slammer | Ordinary Drink | Highball glass | [Summer] | Pour all ingredients (except for lemon juice) ... | [{'id': 18, 'name': 'Amaretto', 'description':... |
| 11 | After Supper Cocktail | Ordinary Drink | Cocktail glass | None | Shake all ingredients with ice, strain into a ... | [{'id': 32, 'name': 'Apricot Brandy', 'descrip... |
| 24 | Amaretto Rose | Ordinary Drink | Collins glass | None | Pour amaretto and lime juice over ice in a col... | [{'id': 18, 'name': 'Amaretto', 'description':... |
| 41 | Blue Lagoon | Ordinary Drink | Highball glass | None | Pour vodka and curacao over ice in a highball ... | [{'id': 1, 'name': 'Vodka', 'description': 'Vo... |
| 13 | Alaska Cocktail | Ordinary Drink | Cocktail glass | [Beach, Chilli] | Stir all ingredients with ice, strain contents... | [{'id': 2, 'name': 'Gin', 'description': 'Gin ... |
| 43 | Blue Mountain | Ordinary Drink | Old-fashioned glass | None | In a shaker half-filled with ice cubes, combin... | [{'id': 1, 'name': 'Vodka', 'description': 'Vo... |
| 77 | Frozen Mint Daiquiri | Ordinary Drink | Old-fashioned glass | None | Combine all ingredients with 1 cup of crushed ... | [{'id': 305, 'name': 'Light Rum', 'description... |
| 4 | Whiskey Sour | Ordinary Drink | Old-fashioned glass | [IBA, Classic, Alcoholic, ContemporaryClassic] | Shake with ice. Strain into chilled glass, gar... | [{'id': 409, 'name': 'Powdered Sugar', 'descri... |
| 0 | Mojito | Cocktail | Highball glass | [IBA, ContemporaryClassic, Alcoholic, USA, Asi... | Muddle mint leaves with sugar and lime juice. ... | [{'id': 170, 'name': 'Soda water', 'descriptio... |

Rysunek 1: Tabela koktajli

# 3 Preprocesing i augmentacja

## 3.1 Tabela składników

W tym punkcie zostanie opracowana tabela wszystkich składników spotykanych we wszystkich koktajlach zbioru. Tabela od razu po wyciągnięciu danych o składnikach z tabeli koktajlów:

| id | name | description | alcohol | type | percentage | imageUrl |
|---|---|---|---|---|---|---|
| 18 | Amaretto | Amaretto (Italian for "a little bitter") is a ... | 1.0 | Liqueur | 28 | https://cocktails.solvro.pl/images/ingredients... |
| 20 | Angostura Bitters | Angostura bitters (English: /æŋgəˈstjʊərə/) is... | 0.0 | Bitter | None | https://cocktails.solvro.pl/images/ingredients... |
| 26 | Apple Brandy | None | 1.0 | Brandy | 35 | https://cocktails.solvro.pl/images/ingredients... |
| 31 | Applejack | Applejack is a strong apple-flavored alcoholic... | 1.0 | Beverage | 40 | https://cocktails.solvro.pl/images/ingredients... |
| 32 | Apricot Brandy | None | 1.0 | Brandy | 24 | https://cocktails.solvro.pl/images/ingredients... |
| ... | ... | ... | ... | ... | ... | ... |
| 520 | White Creme de Menthe | Crème de menthe (pronounced [kʁɛm də mɑ̃t], Fr... | 1.0 | Liquer | None | https://cocktails.solvro.pl/images/ingredients... |
| 528 | Wine | Wine (from Latin vinum) is an alcoholic bevera... | 1.0 | Wine | 14 | https://cocktails.solvro.pl/images/ingredients... |
| 529 | Worcestershire Sauce | Worcestershire sauce (/ˈwʊstərʃər/ (About this... | 0.0 | Sauce | None | https://cocktails.solvro.pl/images/ingredients... |
| 532 | Yellow Chartreuse | Chartreuse (pronounced [ʃaʁtʁøz]) is a French ... | 1.0 | Liqueur | None | https://cocktails.solvro.pl/images/ingredients... |
| 299 | lemon | The lemon, Citrus limon (L.) Osbeck, is a spec... | 0.0 | Fruit | None | https://cocktails.solvro.pl/images/ingredients... |

102 rows × 6 columns

Rysunek 2: Tabela składników

### 3.1.1 Typy składników

Widzimy że tabela zawiera 102 składniki i 6 kolumn dla ich opisania Dalej widać że w tabeli jest kolumna **type**, w której widzimy podobne znaczenia:

- Liqueur i Liquer
- Bitter i Bitters
- A oznaczony jako Beverage Applejack, tak naprawdę jest typu Brandy

```
ingredients['type'].unique()
```

```
array(['Liqueur', 'Bitter', 'Brandy', 'Beverage', 'Rum', None, 'Whiskey',
       'Spirit', 'Liquer', 'Water', 'Wine', 'Cream', 'Soft Drink',
       'Fortified Wine', 'Gin', 'Juice', 'Syrup', 'Soda', 'Fruit',
       'Vodka', 'Flower', 'Bitters', 'Mineral', 'Whisky', 'Sauce', 'Tea'],
      dtype=object)
```

Rysunek 3: Typy składników

Dalej uogólnimy type składników dla dalszej analizy za pomocą podanego mapper'a

```
# Creating less specific types of ingridients for future analysis
def ingredient_type_mapper(ingr_type):
  ingredient_mapping = {
      'Liqueur': 'Alcoholic',
      'Bitter': 'Alcoholic',
      'Brandy': 'Alcoholic',
      'Rum': 'Alcoholic',
      'Whiskey': 'Alcoholic',
      'Whisky': 'Alcoholic',
      'Spirit': 'Alcoholic',
      'Wine': 'Alcoholic',
      'Fortified Wine': 'Alcoholic',
```

```
        'Gin': 'Alcoholic',
        'Vodka': 'Alcoholic',
        'Water': 'Non-Alcoholic',
        'Soft Drink': 'Non-Alcoholic',
        'Juice': 'Non-Alcoholic',
        'Syrup': 'Non-Alcoholic',
        'Soda': 'Non-Alcoholic',
        'Tea': 'Non-Alcoholic',
        'Cream': 'Toppings',
        'Sauce': 'Toppings',
        'Mineral': 'Toppings',
        'Fruit': 'Fruit',
        'Flower': 'Decoration'
    }

    return ingredient_mapping.get(ingr_type, pd.NA)

ingredients['generalized_type'] = ingredients['type'].apply(ingredient_type_mapper)
```

Takie podejście popełnia błędy w raziee jeżeli alkoholowy składnik nie ma typu, ale później to będzie naprawione.

### 3.1.2 Mocność alkolowych składników

Jeżeli sprawdzić jakie typy alkoholi mają choćby jeden wpis z danymi o mocności, to można zauważyć że takich typów składników jest 8, a wszystkich typów alkoholi jest 11. Zakładając że te same typy alkoholi mają podobną mocność, możemy uzupełnić dane o mocności 8 typów alkoholi (tak naprwadę dla 9, bo w tabeli Whisky i Whiskey to są różne typy, chociaż mają podobną mocność):

```
ingredients[ingredients['percentage'].notna()]['type'].unique()
ingredients[ingredients['percentage'].notna()]
# We see that pretty much every type that we can meet in table has at least one ingredient with filled percentage,
# which means that we can fill percentage of other ingredients of the same type, making an assumption that resulting percentage will be mean value of percentage of ingredients of the same type
```

| id | name | description | alcohol | type | percentage | imageUrl | generalized_type |
|---|---|---|---|---|---|---|---|
| 18 | Amaretto | Amaretto (Italian for "a little bitter") is a ... | 1.0 | Liqueur | 28 | https://cocktails.solvro.pl/images/ingredients... | Alcoholic |
| 26 | Apple Brandy | None | 1.0 | Brandy | 35 | https://cocktails.solvro.pl/images/ingredients... | Alcoholic |
| 31 | Applejack | Applejack is a strong apple-flavored alcoholic... | 1.0 | Brandy | 40 | https://cocktails.solvro.pl/images/ingredients... | Alcoholic |
| 32 | Apricot Brandy | None | 1.0 | Brandy | 24 | https://cocktails.solvro.pl/images/ingredients... | Alcoholic |
| 37 | Añejo Rum | Rum is a distilled alcoholic beverage made fro... | 1.0 | Rum | 38 | https://cocktails.solvro.pl/images/ingredients... | Alcoholic |
| 66 | Blended Whiskey | None | 1.0 | Whiskey | 40 | https://cocktails.solvro.pl/images/ingredients... | Alcoholic |
| 97 | Champagne | Champagne (French: [ʃɑ̃.paɲ]) is a sparkling w... | 1.0 | Wine | 13 | https://cocktails.solvro.pl/images/ingredients... | Alcoholic |
| 179 | Dark Rum | None | 1.0 | Rum | 40 | https://cocktails.solvro.pl/images/ingredients... | Alcoholic |
| 2 | Gin | Gin is a distilled alcoholic drink that derive... | 1.0 | Gin | 40 | https://cocktails.solvro.pl/images/ingredients... | Alcoholic |
| 425 | Red Wine | Wine (from Latin vinum) is an alcoholic bevera... | 1.0 | Wine | 14 | https://cocktails.solvro.pl/images/ingredients... | Alcoholic |
| 4 | Tequila | Tequila (Spanish pronunciation: [teˈkila] (Abo... | 1.0 | Spirit | 40 | https://cocktails.solvro.pl/images/ingredients... | Alcoholic |
| 1 | Vodka | Vodka is a distilled beverage composed primari... | 1.0 | Vodka | 40 | https://cocktails.solvro.pl/images/ingredients... | Alcoholic |
| 528 | Wine | Wine (from Latin vinum) is an alcoholic bevera... | 1.0 | Wine | 14 | https://cocktails.solvro.pl/images/ingredients... | Alcoholic |

Rysunek 4: Typy alkoholi mające dane o mocności

Wcześniej tylko 13 alkoholywch składników mieli dane o mocności, po uzupełnieniu już 44 z 50 wszystkich alkoholowych składników.

## 3.2 Tabela koktajli i składników

Mając osobne tabeli dla koktajli i składników, potrzebujemy jednej któraby ich łączyła. Dlatego z tabeli cocktails która nadal zawiera kolumnę ze składnikami dla koktejla, wyciągamy dane o składnikach do innej tabeli:

cocktails_and_ingredients

| | cocktail_id | cocktail_name | ingredient_id | ingredient_name | measure |
|---|---|---|---|---|---|
| 0 | 0 | Mojito | 170 | Soda water | None |
| 1 | 0 | Mojito | 305 | Light Rum | 2-3 oz |
| 2 | 0 | Mojito | 312 | Lime | Juice of 1 |
| 3 | 0 | Mojito | 337 | Mint | 2-4 |
| 4 | 0 | Mojito | 476 | Sugar | 2 tsp |
| ... | ... | ... | ... | ... | ... |
| 526 | 132 | Queen Elizabeth | 189 | Dry Vermouth | 1/2 oz |
| 527 | 133 | Quentin | 179 | Dark Rum | 1 1/2 oz |
| 528 | 133 | Quentin | 282 | Kahlua | 1/2 oz |
| 529 | 133 | Quentin | 304 | Light Cream | 1 oz |
| 530 | 133 | Quentin | 344 | Nutmeg | 1/8 tsp grated |

531 rows × 5 columns

Rysunek 5: Tabela koktajli i składniki

Widać że ona zawiera 531 wiersza i 5 kolumn, z których **measure** nas najbardziej interesuje. Spośród wszystkich wierszy tylko 35 nie mają danych o ilości składnika w koktajlu.

### 3.2.1 Parsing ilości

Nas interesują tylko ilości które można skonwertować w uncji(dalej **oz**), ponieważ tylko oni będą wpływać na mocność koktajla(dalej **ABV**). Konwertacji w oz poddają się kolejny ilości:

- oz
- tblsp - łyżka stołowa
- tsp - łyżka herbatna
- Juice of - sok (dotyczy limonek i cytryn)

```
cocktails_and_ingredients[cocktails_and_ingredients['measure'].str.contains('oz', na=False)]['measure'].unique()
```

```
array(['2-3 oz ', '1/2 oz ', '1 oz ', '2 oz ', '1/3 oz ', '1 2/3 oz ',
       '1 1/2 oz ', '2 1/2 oz Blended ', '3/4 oz ', '8 oz ',
       '1/2 oz white ', '1 oz white ', '6 oz hot ', '1 oz Green Ginger ',
       '3 oz ', '2 1/2 oz ', '1 1/4 oz ', '5 oz ', '1/2 oz Muscatel ',
       '4 oz ', '3/4 oz white ', '1/3 oz cream ', '1/2 oz cream '],
      dtype=object)
```

```
cocktails_and_ingredients[cocktails_and_ingredients['measure'].str.contains("tblsp|tsp", case=False, na=False)]['measure'].unique()
```

```
array(['2 tsp ', '1/2 tsp ', '1 tsp ', '1 tblsp ', '1 1/2 tsp ',
       '1 1/4 tsp ', '1 tsp superfine ', '1/2 tsp superfine ',
       '1/8 tsp grated ', '2 tsp', '1/4 tsp', '1 tblsp', '1/4 tsp '],
      dtype=object)
```

```
cocktails_and_ingredients[cocktails_and_ingredients['measure'].str.contains("juice", case=False, na=False)]['measure'].unique()
```

```
array(['Juice of 1 ', 'Juice of 1/2 ', 'Juice of 1/2', 'Juice of 1/4 '],
      dtype=object)
```

Rysunek 6: Ilości do parsingu

Po parsingu wszystkich ilości możemy zobaczyć wyniki na przykładzie z Mojito, które zawiera składniki ze wszystkimi wymienionymi ilościami oprócz tblsp:

| | cocktail_id | cocktail_name | ingredient_id | ingredient_name | measure | volume_oz |
|---|---|---|---|---|---|---|
| 0 | 0 | Mojito | 170 | Soda water | None | NaN |
| 1 | 0 | Mojito | 305 | Light Rum | 2-3 oz | 2.50 |
| 2 | 0 | Mojito | 312 | Lime | Juice of 1 | 1.01 |
| 3 | 0 | Mojito | 337 | Mint | 2-4 | NaN |
| 4 | 0 | Mojito | 476 | Sugar | 2 tsp | 0.27 |

Rysunek 7: Wynik parsinga

## 3.3 Tabela koktajli

### 3.3.1 Obliczanie mocności koktajli

Przed tym jak przejdziemy do obliczania ABV każdego koktajla, zwrócę uwagę na to że na razie to jest możliwe tylko z popełnieniem błędów, albo z pominięciem dużej ilości koktajlów.

```
# For each cocktail, calculate its approximate ABV
for cocktail_id, group in result_df.groupby('cocktail_name'):
    essential_ingrs = []
    for index, row in group.iterrows():

      gen_type = row['generalized_type'] if pd.notna(row['generalized_type']) else "Unknown"
      volume_oz = row['volume_oz']
      percentage = row['percentage']

      if pd.notna(percentage) and pd.notna(volume_oz) and gen_type in ('Alcoholic', 'Non-
          Alcoholic', 'Fruit'):
        essential_ingrs.append([row['percentage'], row['volume_oz']])

    total_volume, total_alcohol_volume, abv = 0, 0, 0

    for percentage, volume in essential_ingrs:
      total_volume += volume
      total_alcohol_volume += (percentage / 100) * volume

    if total_volume > 0 and total_alcohol_volume > 0:
      abv = (total_alcohol_volume / total_volume) * 100
    else:
      abv = None

    cocktails.loc[cocktails['name'] == cocktail_id, 'abv'] = abv
```

Podany kod obliczy ABV dla wszystkich koktajlów które zawierają choćby jeden składnik typu Alcoholic, Non-Alcoholic lub Fruit, który ma dane o jego ilości w oz w tym koktajlu. Takie podejście jest nieprawidłowe i doprowadza do tego że 128 koktajlów będą mieli dane o ABV, ale tylko 54 z nich będą mieli napewno prawdziwe wartości, bo oni nie zawierają składników bez **generalized-type**. Bo jeżeli koktajl zawiera składnik bez **generalized-type** i **volume-oz**, który potencjalnie może być typu **Non-Alcoholic**, algorytm obliczy ABV bez uwzględnienia tego składniku, co doprowadzi do błędu.

Aby pozbyć się tego problemu, popatrzymy jakie składniki nie mają uogólnienego typu:

```
ingredients.query('generalized_type.isna()')
```

| id | name | description | type | percentage | imageUrl | generalized_type |
|---|---|---|---|---|---|---|
| 47 | Banana | None | None | 0.0 | https://cocktails.solvro.pl/images/ingredients... | <NA> |
| 53 | Benedictine | None | None | 0.0 | https://cocktails.solvro.pl/images/ingredients... | <NA> |
| 56 | Bitters | None | None | 0.0 | https://cocktails.solvro.pl/images/ingredients... | <NA> |
| 94 | Celery Salt | None | None | 0.0 | https://cocktails.solvro.pl/images/ingredients... | <NA> |
| 106 | Cherry | None | None | 0.0 | https://cocktails.solvro.pl/images/ingredients... | <NA> |
| 127 | Club Soda | None | None | 0.0 | https://cocktails.solvro.pl/images/ingredients... | <NA> |
| 194 | Egg | None | None | 0.0 | https://cocktails.solvro.pl/images/ingredients... | <NA> |
| 192 | Egg White | None | None | 0.0 | https://cocktails.solvro.pl/images/ingredients... | <NA> |
| 193 | Egg Yolk | None | None | 0.0 | https://cocktails.solvro.pl/images/ingredients... | <NA> |
| 270 | Ice | None | None | 0.0 | https://cocktails.solvro.pl/images/ingredients... | <NA> |
| 294 | Lemon Peel | None | None | 0.0 | https://cocktails.solvro.pl/images/ingredients... | <NA> |
| 296 | Lemon vodka | None | None | 0.0 | https://cocktails.solvro.pl/images/ingredients... | <NA> |
| 322 | Maraschino Cherry | None | None | 0.0 | https://cocktails.solvro.pl/images/ingredients... | <NA> |
| 344 | Nutmeg | None | None | 0.0 | https://cocktails.solvro.pl/images/ingredients... | <NA> |
| 347 | Olive | None | None | 0.0 | https://cocktails.solvro.pl/images/ingredients... | <NA> |
| 409 | Powdered Sugar | None | None | 0.0 | https://cocktails.solvro.pl/images/ingredients... | <NA> |
| 170 | Soda water | None | None | 0.0 | None | <NA> |
| 476 | Sugar | Sugar is the generic name for sweet-tasting, s... | None | 0.0 | https://cocktails.solvro.pl/images/ingredients... | <NA> |

Rysunek 8: Składniki bez generalized-type

Takich które owszem wpływają na ABV koktajlu jest niedużo, mianowice:

- Benedictine - jest mocnym likierem
- Bitters - rodzina mocnych likierów
- Club Soda
- Lemon vodka
- Soda water

I od razu na miejscu naprawimy te typy które nie uwzględnił mapper w punkcie 3.1.1.

Znów powtarzamy obliczenie ABV, ale już innym algorytmem który uwzględnia brak danych o ilości składnika który jest typu Alcoholic, Non-Alcoholic. Teraz już 68 koktajli mają dane o mocności, co oznacza że naprawienie typów niektórych składników i zmiana algorytmu uratowali nam dane o mocności 14 koktajli.

```
cocktails[cocktails['abv'].notna()]
```

| | name | category | glass | tags | instructions | abv |
|---|---|---|---|---|---|---|
| 4 | Whiskey Sour | Ordinary Drink | Old-fashioned glass | [IBA, Classic, Alcoholic, ContemporaryClassic] | Shake with ice. Strain into chilled glass, gar... | 40.000000 |
| 6 | Daiquiri | Ordinary Drink | Cocktail glass | [IBA, Classic, Beach] | Pour all ingredients into shaker with ice cube... | 29.177057 |
| 7 | Margarita | Ordinary Drink | Cocktail glass | [IBA, ContemporaryClassic] | Rub the rim of the glass with the lime slice t... | 24.666667 |
| 9 | Moscow Mule | Punch / Party Drink | Copper Mug | [IBA, ContemporaryClassic] | Combine vodka and ginger beer in a highball gl... | 6.666667 |
| 10 | After Dinner Cocktail | Ordinary Drink | Cocktail glass | [DinnerParty] | Shake all ingredients (except lime wedge) with... | 26.000000 |
| ... | ... | ... | ... | ... | ... | ... |
| 123 | Pink Lady | Ordinary Drink | Cocktail glass | None | Shake all ingredients with ice, strain into a ... | 36.697248 |
| 124 | Poppy Cocktail | Ordinary Drink | Cocktail glass | None | Shake ingredients with ice, strain into a cock... | 36.000000 |
| 125 | Port And Starboard | Ordinary Drink | Pousse cafe glass | None | Pour carefully into a pousse-cafe glass, so th... | 14.432990 |
| 128 | Quaker's Cocktail | Ordinary Drink | Cocktail glass | None | Shake all ingredients with ice, strain into a ... | 27.558140 |
| 133 | Quentin | Ordinary Drink | Cocktail glass | None | In a shaker half-filled with ice cubes, combin... | 37.000000 |

68 rows × 6 columns

Rysunek 9: Koktajli z danymi o ABV

### 3.3.2 Kategorię mocności koktajli

Pod koniec rozdzielimy koktajli na kategorię według ich mocności:

- ABV is `na` - Unknown
- ABV $< 10\%$ - Weak
- $10\% \leqslant$ ABV $< 20\%$ - Moderate
- $20\% \leqslant$ ABV $< 30\%$ - Strong
- ABV $\geqslant 30\%$ - Very Strong

### 3.3.3 Długość instrukcji, liczba składników, sposób przyrządzenia

Dla przyszłej analizy dodamy kolumnę z długością instrukcji, liczbą składników w koktajlach i ich sposobem przyrzadzenia(ekstraktowany z instrukcji):

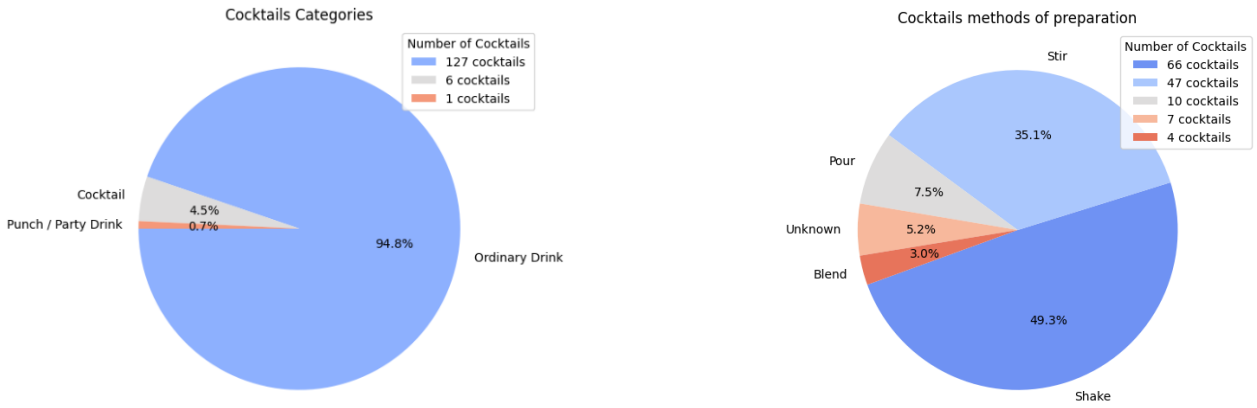
| | name | category | glass | tags | instructions | abv | strength | instruction_length | num_ingredients | prep_method |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Mojito | Cocktail | Highball glass | [IBA, ContemporaryClassic, Alcoholic, USA, Asi... | Muddle mint leaves with sugar and lime juice. ... | NaN | Unknown | 177 | 5 | Unknown |
| 1 | Old Fashioned | Cocktail | Old-fashioned glass | [IBA, Classic, Alcoholic, Expensive, Savory] | Place sugar cube in old fashioned glass and sa... | NaN | Unknown | 218 | 4 | Unknown |
| 2 | Long Island Tea | Ordinary Drink | Highball glass | [Strong, Asia, StrongFlavor, Brunch, Vegetaria... | Combine all ingredients (except cola) and pour... | NaN | Unknown | 152 | 6 | Unknown |
| 3 | Negroni | Ordinary Drink | Old-fashioned glass | [IBA, Classic] | Stir into glass over ice, garnish and serve. | NaN | Unknown | 44 | 3 | Stir |
| 4 | Whiskey Sour | Ordinary Drink | Old-fashioned glass | [IBA, Classic, Alcoholic, ContemporaryClassic] | Shake with ice. Strain into chilled glass, gar... | 40.0 | Very Strong | 148 | 4 | Shake |

Rysunek 10: Koktajli z nowymi kolumnami

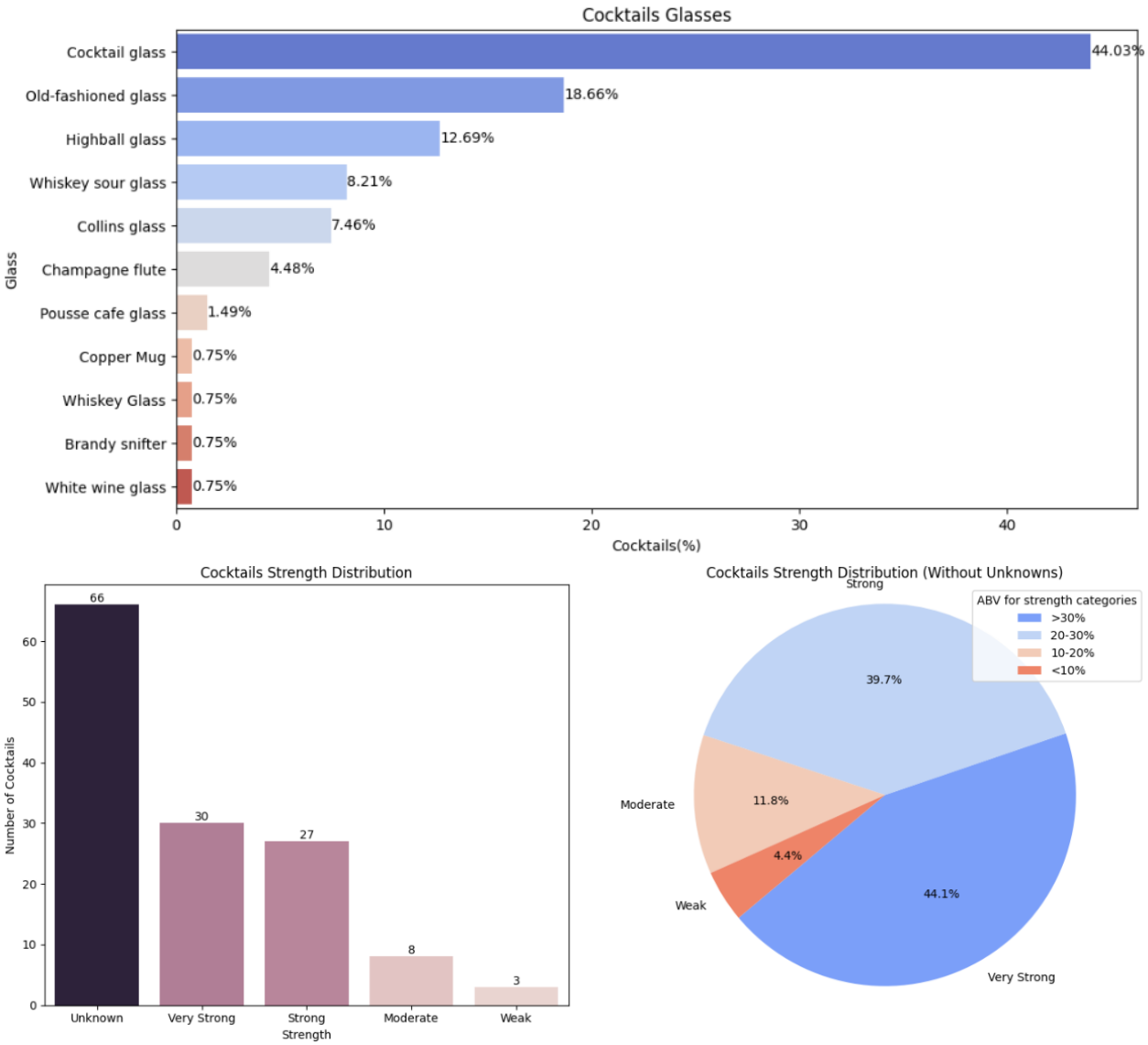
# 4 Ewaluacja

W tym punkcie będą załączone wszelkie wykresy pokazujący charakter naszych danych.

## 4.1 Koktajli

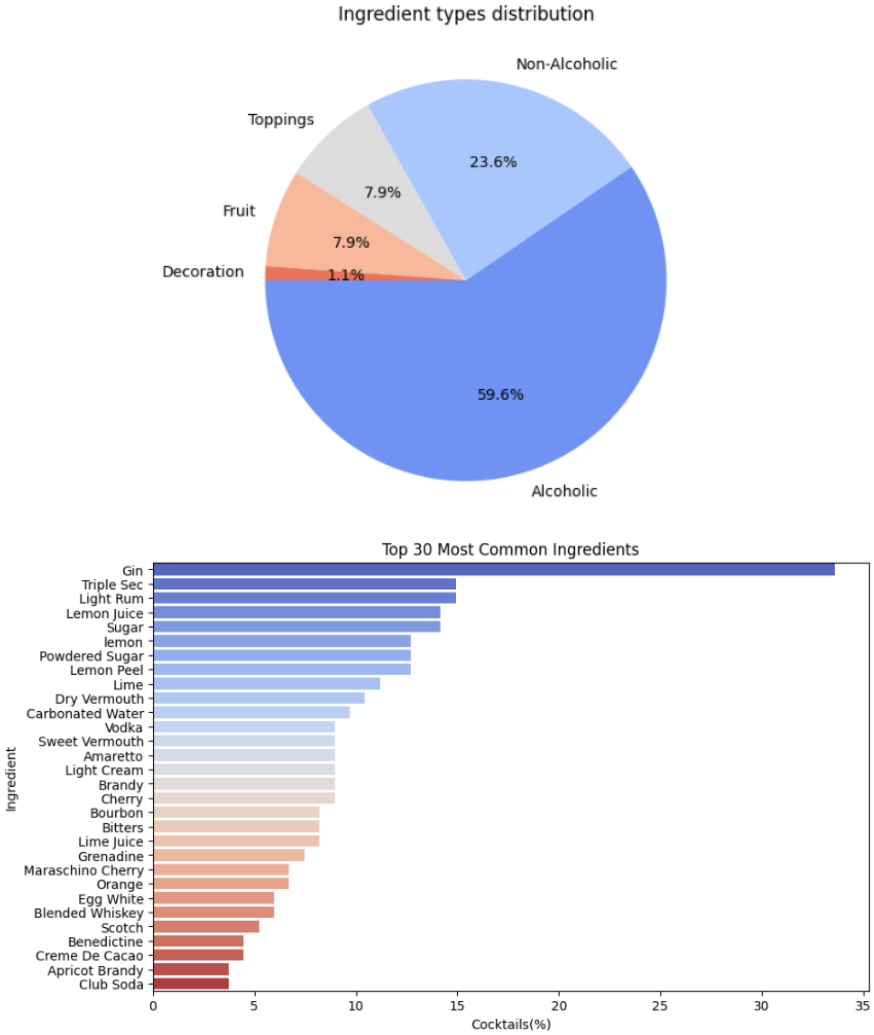



Rysunek 11: Jak widać prawie wszystkie koktajli są Ordinary Drinka'ami, i prawie połowe koktajli trzeba mieszać w szejkerze.

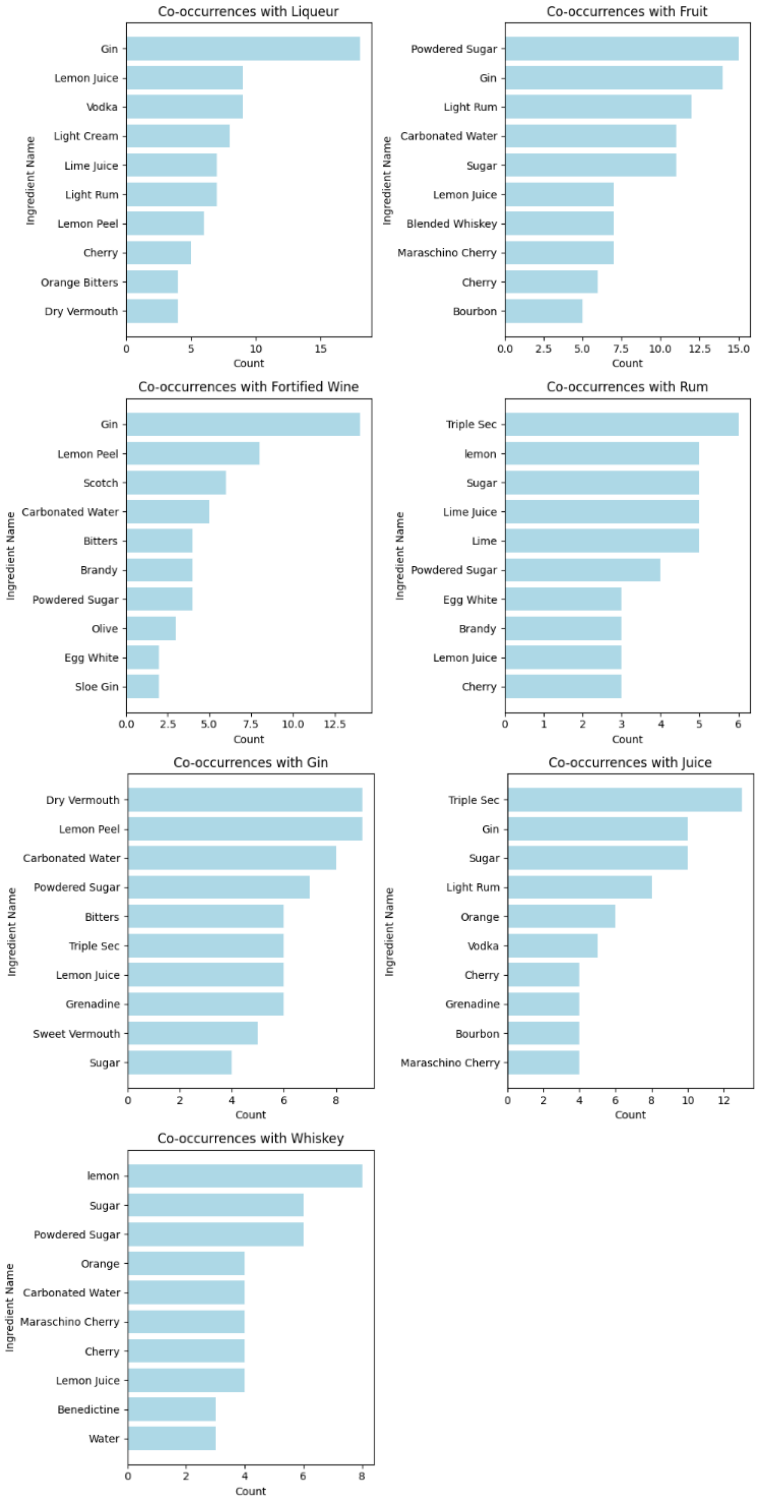



Rysunek 12: Większość koktejli jest nalewana w szklankę koktajlową, a aż 85% koktajlów jest mocniej 20%.

## 4.2 Składniki

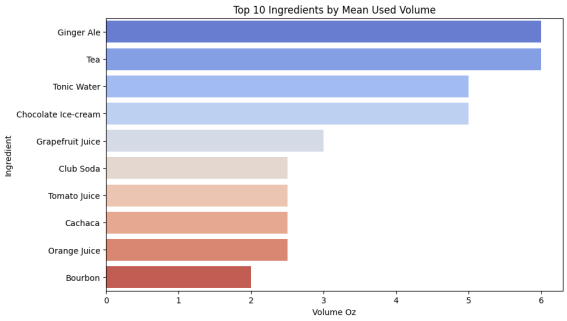



Rysunek 13: Widzimy że ponad połowa składników jest alkoholowa, z których Gin jest wykorzystywany najczęściej. A sok cytryny jak widać jest jednym z najpopularniejszych składników.

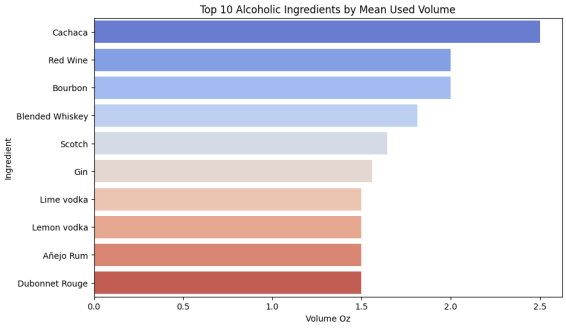



Rysunek 14: Widzimy że najpopularniejszy gin zazwyczaj razem nie występuje tylko z romem i whisky. Ale widać że gin często jest razem z vermutem - jak jeszcze przyrządzić wszelkie martini? Likiery często pijemy z sokiem cytryny. A do owoców często dodajemy cukier. Whiski często serwujemy z cytryną(z sokiem rzadziej). Natomiast rom też serwowany z cytryną lub limonką, zazwyczaj pijemy z sokiem limonki a nie cytryny.

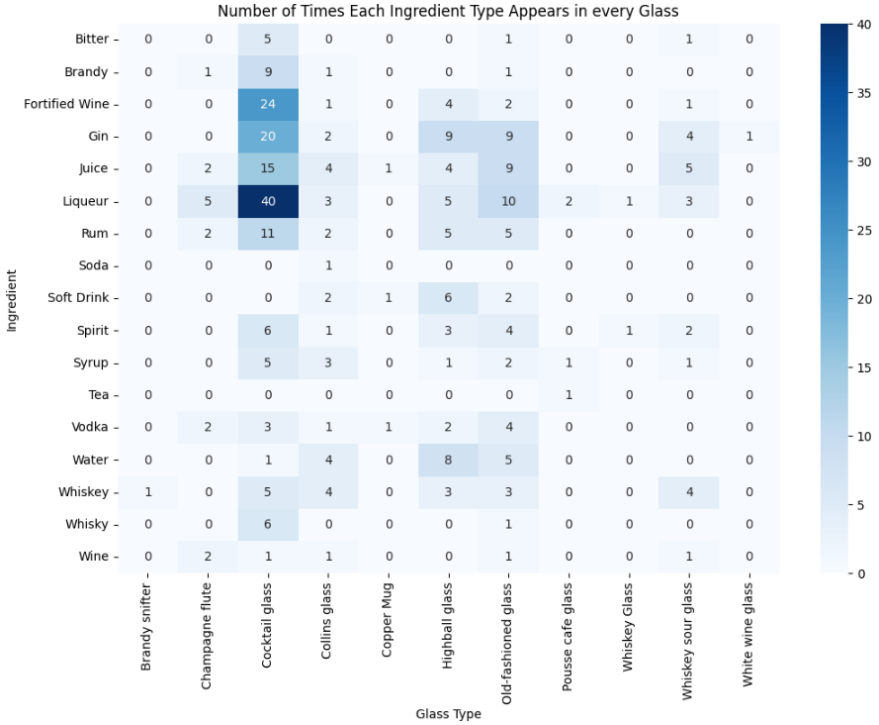
## 4.3 Koktajli i składniki



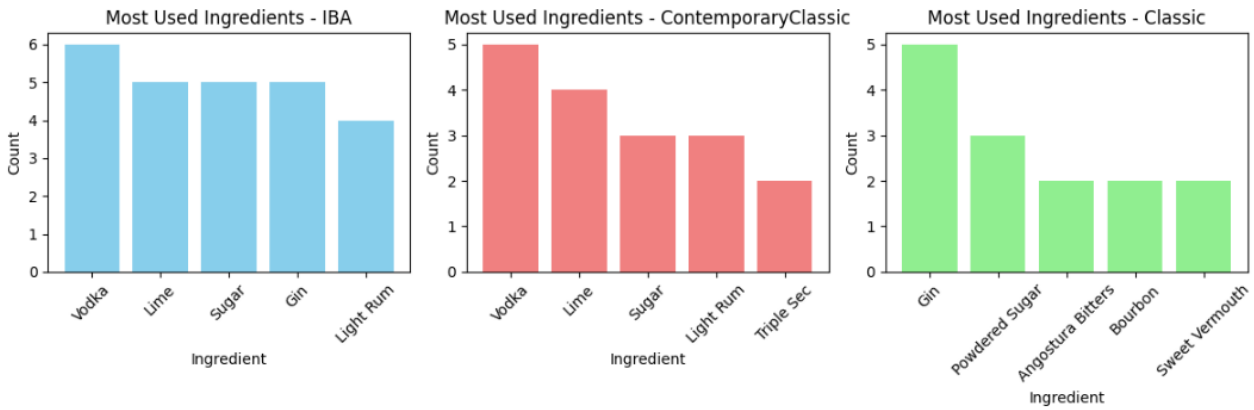

(a) Mocno rozwadniamy wódkę w Moscow Mule ginger ale'em, i herbatą Amaretto Tea - deserowy koktajl.



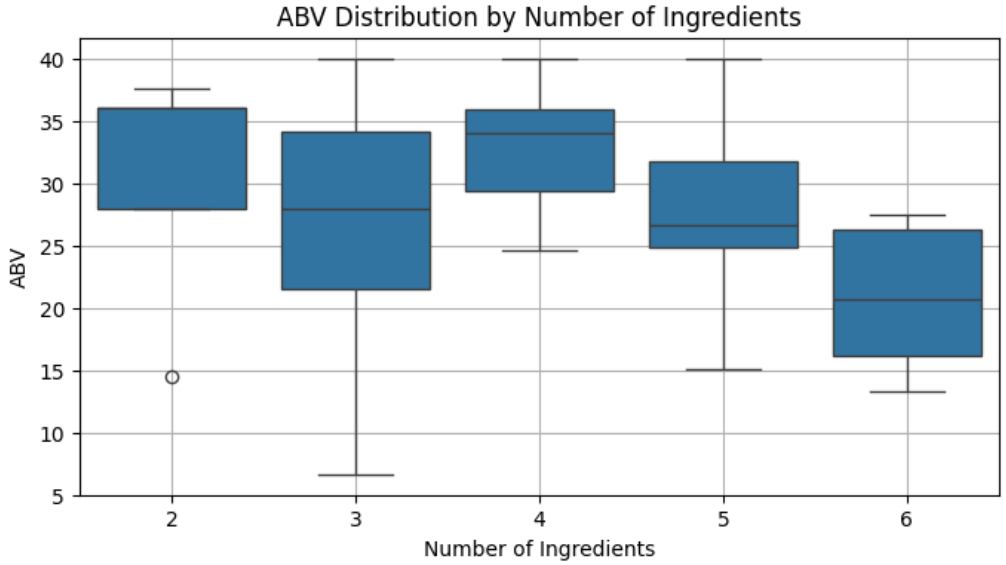

(b) Robimy mocne koktajli z cachaca i bourbonem, i trochę mniej mocne z ginem i wódkami.




Rysunek 16: Jak widać, większość koktajlów jest nalewana szklankę koktajlową - w tym najwięcej koktajlów na podtsawie likierów i wzmocnionych win i gina. Widać że sporo składników jest stosowana w koktajlach w old-fashioned glass.




Rysunek 17: Widzimy, że klasyczne koktajle są zwykle robione z Ginem (Martini itp.), podczas gdy współczesna Klasyka bardziej koncentruje się na wódce z limonką. A koktajle IBA powstają głównie z wódki i ginu.




Rysunek 18: Na tym wykresie widać trend na spadek mocności koktajlów przy zwiększeniu liczby składników. Serwując tylko dwa składniki - chcemy posmakować ich kombinację bez rozwodnienia. Serwując 6 - szukamy nowych smaków na podstawie alkoholowych składników.

# 5 Ciekawostka

Jeżeli nie chcemy wydawać pieniędzy na wycieczkę do baru, aby spróbować ciekawe kombinacji składników odnalezione podczas analizy - możemy kupić kilka składników i przyrządzić koktajlę w domu. Ale które składniki warto kupić jeżeli chcemy przyrządzić jak najwięcej koktajlów?

**optimizer.py** zawiera rozwiązanie tego zadania liniowego programowania, które pomoże określić te **n** składników za pomocą których potrafimy spróbować jak najwięcej koktajlów. Optimizer posiada możliwość rozwiązania tego zadania na dwa sposoby:

- Odnalezienie **n** składników za pomocą których będzie można przyrządzić największą liczbę koktajli
- Odnalezienie **n** alkoholowych składników na podstawie których, z dodaniem reszty nie alkoholowych składników, można przygotować jak najwięcej koktajli

Drugi punkt brzmi ciekawiej, bo zakupienie alkoholowych składników jest dość drogie, co oznacza że nie potrafimy dużo ich kupić, a nie alkoholowe: wszelkie soki, owoce i toniki nie są aż tak drogie, i ich już można kupić dość dużo i przyrządzić tym więcej koktajlów.

Przykład działania:

```
opt = Optimizer(ingredients, cocktails_and_ingredients)
result = opt.find_n_ingredients_to_make_largest_amount_of_cocktails(7)
opt.print_results(result)
```

```
With 7 ingredients, you can make 8 cocktails!

Selected ingredients and their usage:
1. Gin (used in 5 cocktails)
2. Creme De Cacao (used in 3 cocktails)
3. Amaretto (used in 3 cocktails)
4. Light Cream (used in 3 cocktails)
5. Sweet Vermouth (used in 2 cocktails)
6. Bitters (used in 2 cocktails)
7. Triple Sec (used in 1 cocktails)

Cocktails you can make:
- Almond Joy
- Amaretto And Cream
- Artillery
- Flying Dutchman
- Foxy Lady
- Lone Tree Cocktail
- Pink Gin
- Poppy Cocktail
```

(a) Te koktajli raczej będą dość podobne do siebie.

```
result = opt.find_n_ingredients_to_make_largest_amount_of_cocktails(2, True)
opt.print_results(result)
```

```
With 2 ingredients, you can make 18 cocktails!

Selected ingredients and their usage:
1. Gin (used in 12 cocktails)
2. Light Rum (used in 6 cocktails)

Rest of needed ingredients:
Soda Water
Lime
Mint
Sugar
Powdered Sugar
Coca-Cola
Ginger Ale
Lime Juice
Pineapple
Tonic Water
Carbonated Water
Lemon Peel
Orange spiral
Grenadine
Lemon Juice
Maraschino Cherry
Orange
lemon
Orange Peel
Water
Cherry
Strawberries
Pineapple Juice
Egg White
Light Cream

Cocktails you can make:
- Cuba Libre
- Daiquiri
- Dragonfly
- Frozen Mint Daiquiri
- Frozen Pineapple Daiquiri
- Gin And Tonic
- Gin Cooler
- Gin Daisy
- Gin Fizz
- Gin Sling
- Gin Smash
- Gin Sour
- Gin Squirt
- Gin Toddy
- Havana Cocktail
- Lady Love Fizz
- Mojito
- Pink Lady
```

(b) Nawet z 2 alkoholi można przyrządzić sporo koktajli, chociaż też dość podobnych. Warto wtedy kupić jeszcze choćby 2 innych alkohola.

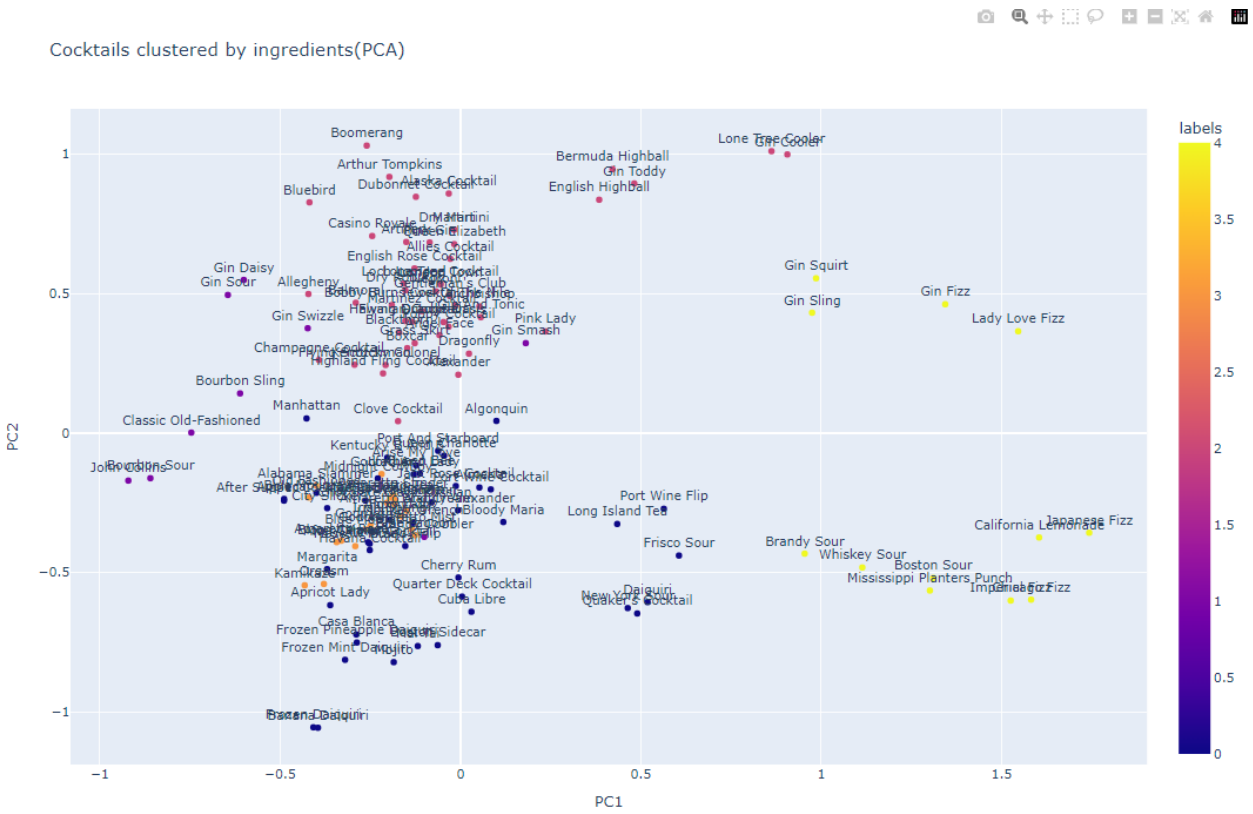
# 6 Klasteryzacja

Głównym celem klasteryzacji koktajli oczywiście jest zgrupowanie ich tak, aby w każdej grupie byli koktajli najbardziej podobne do siebie.
Najsensowniej określać podobieństwo koktajli w zależności od ich składników, od tego i zaczniemy.

## 6.1 Klasteryzacja na podstawie składników

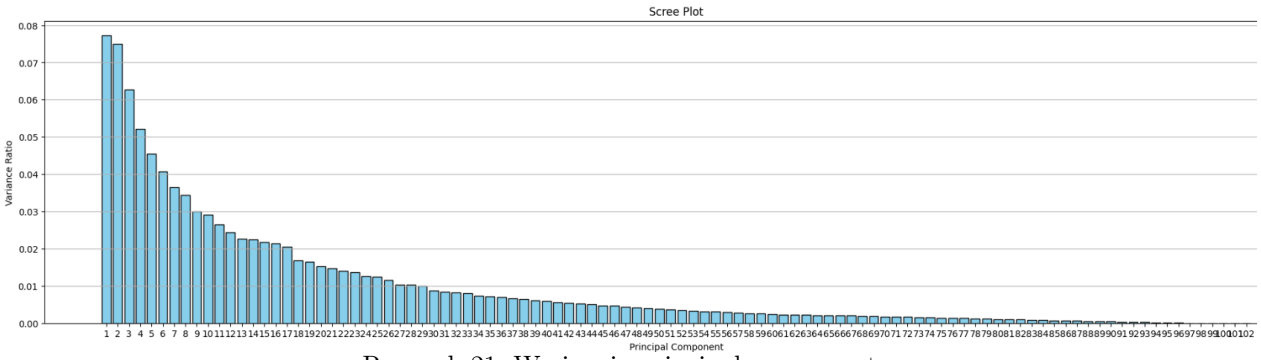
Klasteryzować będziemy na podstawie objętości składników w koktajlach, jeżeli nie mamy danych o objętości w oz pewnych składników w koktajlu, zamieniamy ich na 0.01 oz, nie najlepsze podejście, ale działa dla wszystkich składników, nawet dla wszelkich jagod i owoców stosowanych dla dekaracji koktajli.

Zaczniemy od klasteryzacji za pomocą KMeans z **n_clusters=5**, bo tyle jest podstawowych smaków - salty, spicy, sweet, sour, bitter - jest to naiwnie wierzyć że KMeans rozdzieli koktajli akkurat według smaków, ale niech będzie 5.



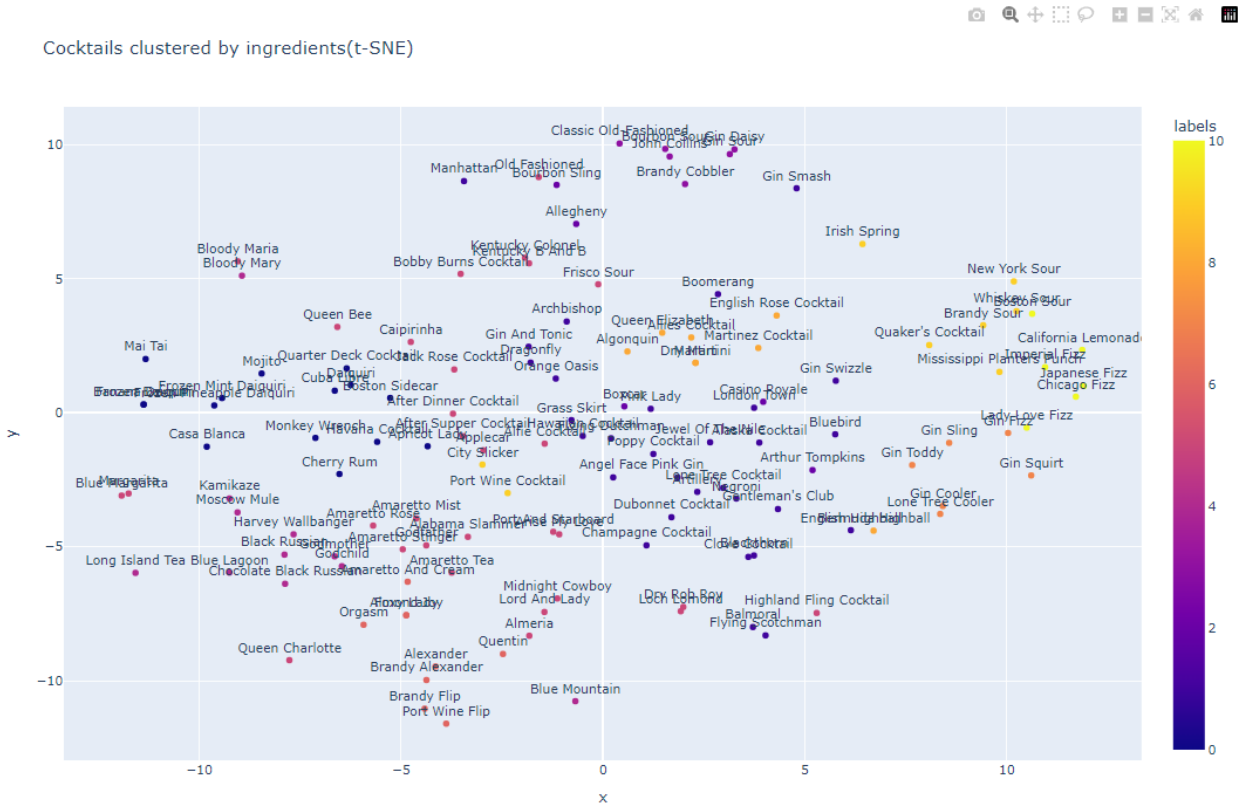

Rysunek 20: Wizualizacja koktajli za pomocą PCA.

Na wykresie widzimy sporo tłumu który bez przybliżenia nie pozwala nawet odczytać nazw koktajlów, ale nawet jeżeli przybliżmy ten wykres, zauważymy że często jest tak że koktajli które są blisko siebie, mogą nawet nie mieć wspólnych składników, co oznacza że PCA nie potrafił dobrze zwizualizować koktajle na 2D wykresie - sprawdzimy dlaczego.



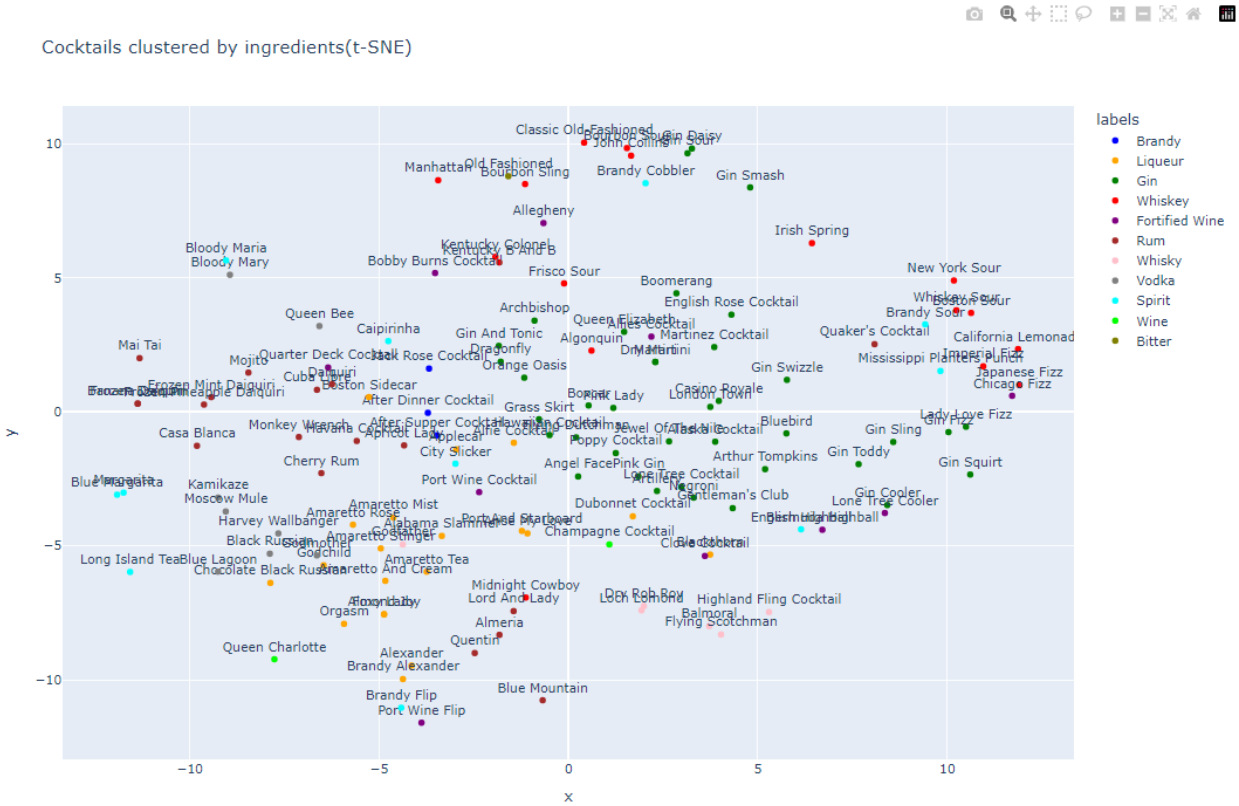

Rysunek 21: Wariancja principal components.

Na tym wykresie widzimy poziomy wariancji wszystkich PC dla naszego PCA. Widać że PC1 i PC2 w sumie odpowiadają tylko za około 16.5% od całej wariancji, co jest za mało dla wizualizacji koktajli na 2D wykresie. Dlatego spróbójmy zastosować inną metodę do wizualizacji.




Rysunek 22: Wizualizacja koktajli za pomocą t-SNE.

Wizualizacja za pomocą bardziej współczesnej metody t-SNE już wygląda o wiele lepiej i nawet jeżeli sprawdzimy składniki koktajli leżących blisko siebie, zauważymy że owszem składniki są podobne!
Jedyna ciekawostka jest w tym, że ten wykres może nieść w sobie jeszcze więcej informacji, bo na razie kolory koktajlów tak naprawdę określają grupy koktajli najbardziej do siebie podobnych, co jest przydatne dla stworzenia pewnego spisu koktajli z rozdzieleniem na grupy według klasterów, ale nic więcej nam nie mówią, i możemy spróbować to zmienić.




Rysunek 23: Wizualizacja koktajli za pomocą t-SNE i kolorowanie na podstawie typu głownego alkoholu.

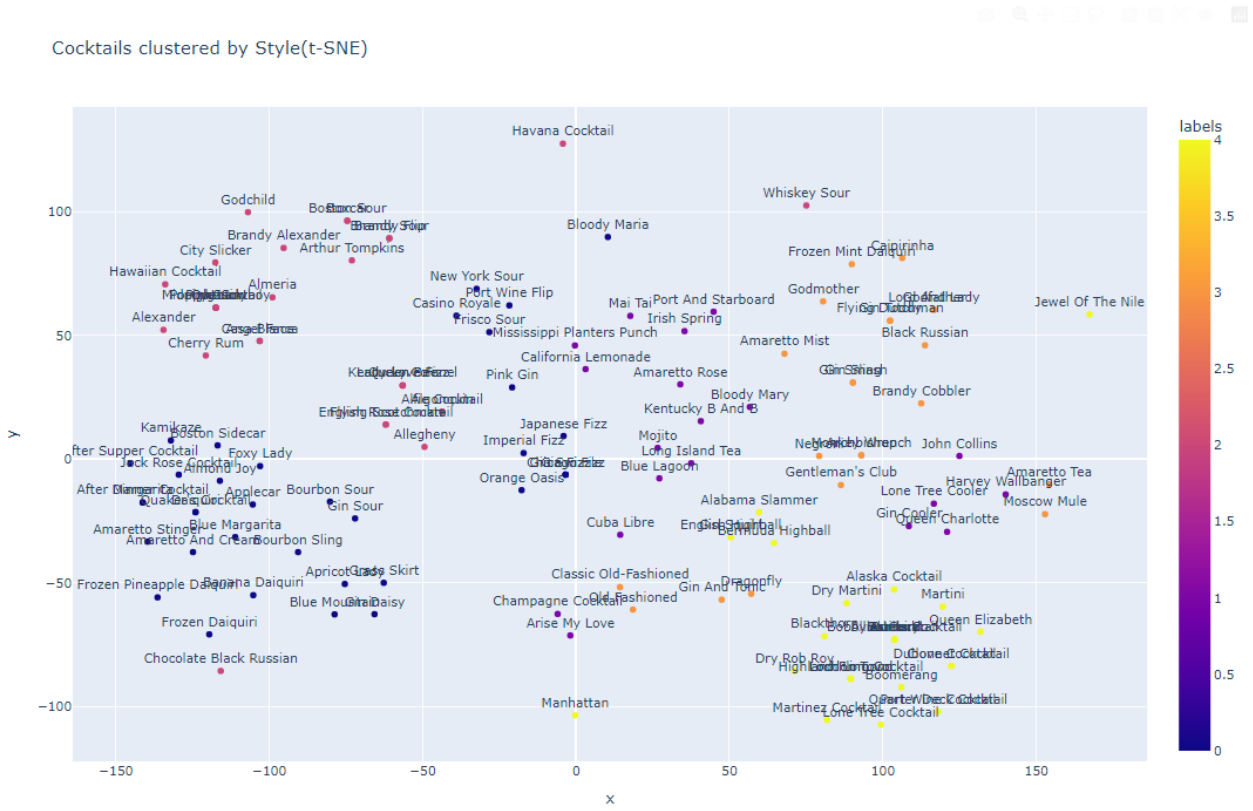
Ten wykres jest jeszcze lepszy od zeszłego, ponieważ teraz kolory mają sens i poza tym nawet widać że zazwyczaj koktajli o tej samej podstawie alkoholowej są blisko siebie, chociaż takie podejście nadal nie jest idealne, bo czasami koktajli mają kilka alkoholowych składników o tej samej objętości i ich głowny jest wybierany ”losowo” spośród tych.

## 6.2 Klasteryzacja na podstawie "stylu" przyrządzenia

Oprócz składników możemy również klasteryzować nasze koktajle na podstawie innych parametrów m.in:

- Glass
- Strength
- Num ingredients
- Preparation method

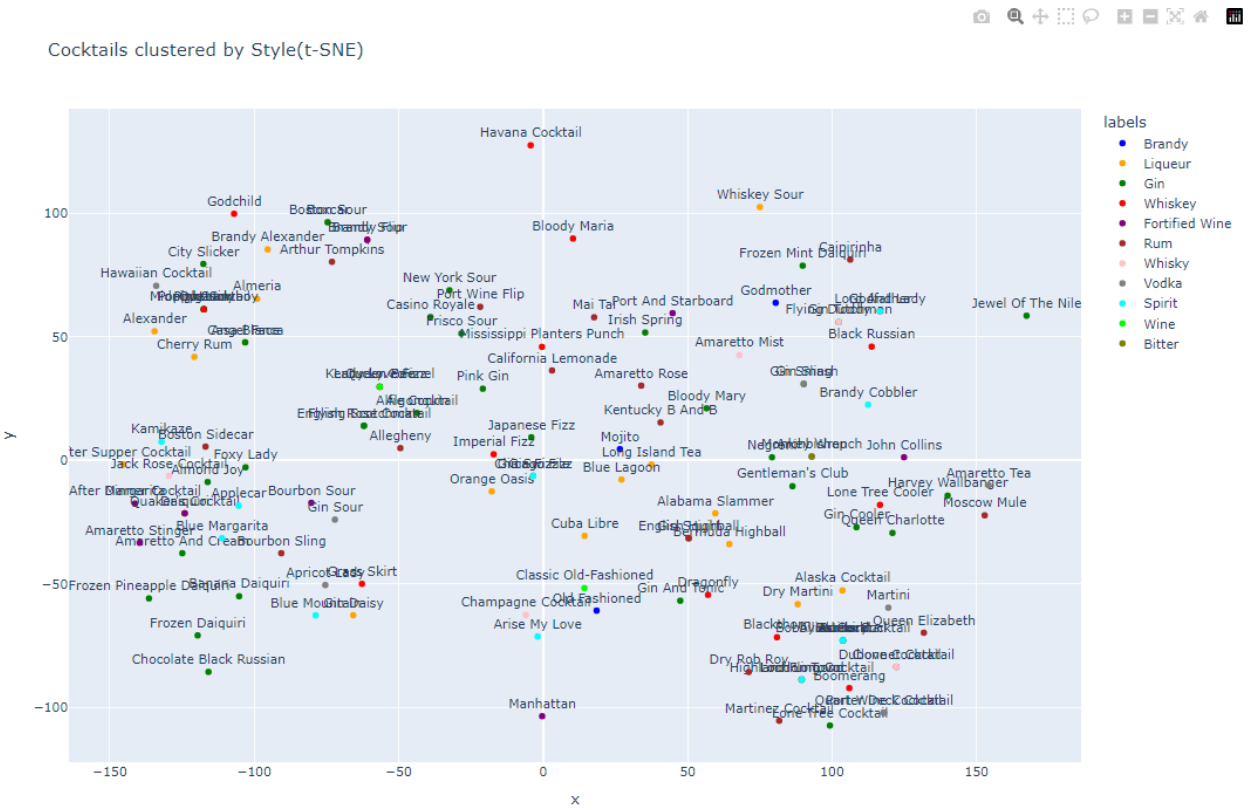
Dla stosowania **Glass** i **Preparation method** w klasteryzacji i wizualizacji musimy przekształcić ich do liczbowej postaci - zrobimy to za pomocą **OneHotEncoder**.




Rysunek 24: Wizualizacja koktajli za pomocą t-SNE i kolorowanie na podstawie KMeans.

Jeżeli porównać koktajli leżące blisko siebie, zauważymy że zazwyczaj są podobnej mocności, przyrządzane tym samym sposobem i mają mniej więcej tyle same składników, ale już rzadziej są nalewane w te same szklanki.

Z ciekawości pomalujemy koktajle na tym wykresie na podstawie typu ich głównego alkoholego składniku.




Rysunek 25: Wizualizacja koktajli za pomocą t-SNE i kolorowanie na podstawie typu głównego składniku.

Widać że koktajle o tych samych kolorach prawie nie formują klastery, z czego wynika że "styl" koktajlu nie bardzo zależy od jego głównego składniku.

# 7 Wnioski

Podczas tej analizy zauważyliśmy sporo zależności między różnymi danymi o koktajlach, ale najcenniejszymi są:

- Gin jest najpopularniejszym składnikom koktajli.
- ABV koktajlów zazwyczaj spada przy zwiększeniu liczby składników.
- Nawet z niedużej ilości różnych alkoholi można przyrządzić wiele koktajlów.
- Da się klasteryzować koktajli i nawet w na tyle dobry sposób, że z tego można korzystać w praktyce.