# Description

Ce projet de nature académique a pour but de nous aider à développer nos compétences techniques dans l’utilisation du langage de programmation R. Le logiciel R studio est utilisé afin de traiter, analyser et présenter des données dans le but de tirer de déceler des tendances et tirer des conclusions. Dans notre cas, un jeu de données comprenant des informations sur les écrasements d’avions qui ont eu lieu entre 1908 et 2009 est utilisé.

# Pourquoi avons-nous choisi ce jeu de données ?

Nos critères principaux pour ce projet étaient d’analyser un jeu de données qui contenait peu de textes, qui ne traitait pas de données économiques et qui traitait d’un sujet original. On entend par original, un sujet qui n’avait pas trop été analysé dans le passé et dont peu de conclusions ont été tirées. En faisant plusieurs recherches, nous sommes tombés sur le site web [www.kaggle.com](http://www.kaggle.com) qui contenait plusieurs jeux de données intéressants sur des sujets diverses. Le jeu de données sur les crashs d’avions nous a paru intéressant puisqu’il répondait à tous nos critères en plus d’être riche en information.

# Description du jeu de données

Le jeu de données couvre 5268 crashs d’avions qui ont eu lieu entre 1908 et 2009. Il inclut les informations suivantes affichées à travers 13 colonnes:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Date | Date du crash |
| Time | Heure du crash |
| Location | Endroit du crash |
| Operator | Opérateur de l'avion |
| Flight.. | Numéro de vol |
| Route | Itinéraire prévu |
| Type | Modèle d’avion |
| Registration | Numéro de registration de l’avion |
| Cn.In | Aucune description fournie |
| Aboard | Nombre de passagers à bord |
| Fatalities | Nombre de morts |
| Ground |  |
| Summary | Détails du crash. Les données sous forme de textuelle |

## Données Manquantes

Il est important de noter que notre jeu de données possède quelques lacunes:

\*\*\* la prof veut savoir combien de données manquantes on a dans notre fichier

1. Pour des raisons inconnues, notre fichier de données n’a enregistré aucun crash entre 1908 et 1912, mis à part un crash en 1908.
2. Les informations sur les crashs survenus au cours des conflits armés semblent aussi manquantes. En effet, on s’attendrait par exemple à voir un pic du nombre d’écrasement militaires au cours de la période allant de 1939 à 1945 (seconde guerre mondiale). Cependant le jeu de données nous produit en moyenne 50 crashs par an au cours de cette période. (Ce n’est pas vrai du tout) --- a retirer—dans mon graphique on voit un boom entre 39 et 45.
3. Certaines variables sont manquantes pour plusieurs des crashs répertoriés dans le jeu de données.
4. Plusieurs variables ont des erreurs d’orthographe, dû au fait que la saisie des données a été faites manuellement
5. Les noms de pays étaient mal écrits dans beaucoup de situation. Il fallait donc normaliser et corriger ces erreurs avant de pouvoir traiter les données.

# Préparation et transformation des données

Une étape de préparation a été nécessaire afin de pouvoir répondre aux questions de l’exercice :

1. Ajout du Colonne représentant la catégorie du vol (Militaire ou Commercial). ???
2. Ajout d’une colonne représentant le pays ou l’écrasement est survenu.
3. Ajout d’une colonne représentant la surface de l’écrasement (Terre ou Mer).
4. Ajout de plusieurs colonnes représentant la destination du vol (ville, pays…)
5. Ajout d’une colonne représentant la raison du vol (entrainement, démonstration, ou non).
6. Ajout d’une colonne représentant la probabilité de survie.
7. Correction de plusieurs erreurs d’orthographe dans le nom des pays.
8. Remplacement des provinces canadiennes par le nom du pays
9. Remplacement des états américains par le nom du pays.
10. Regroupement par continent ou par hémisphère pour avoir des taxonomies niveau continent.

# Packages Utilisés – ce n’est pas nécessaire

Les packages non inclus dans R base que nous avons utilisé pour ce devoir sont :

* ggmap
* plyr
* data.table

# Questions et Réponses

## « Le nombre de crashs à travers le temps »

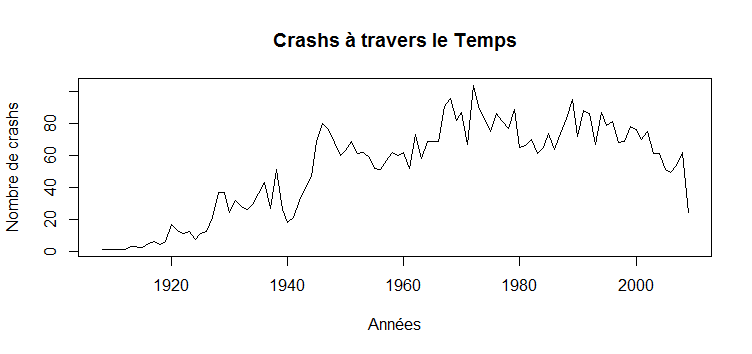
### Description

A rajouter…

### Approche de résolution

* Extraction de l’année (les 4 derniers caractères) de la variable « Date » à l’aide de la fonction substr et application de ce vecteur a la variable à « Annee »
* Combination du vecteur appelé « Annee » et d’une partie du jeu de données initial a « table\_temporaire » à l’aide de la fonction cbind.
* Appel de la fonction « trouver\_frequence » et application du resultat à la variable « x »
* Création d’un data.frame appelé « table\_occurence\_annee » avec « x » (vecteur de fréquence des dates) et le vecteur d’années sans duplicats
* Nomination des colonnes du data.frame
* Ordonne « table\_occurence\_annee » en ordre croissant par date

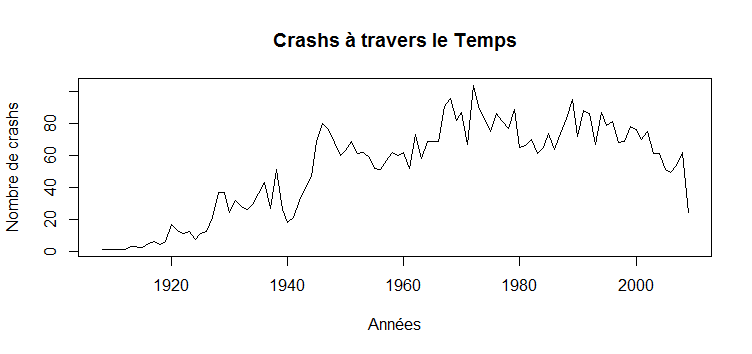
### Résultats et interprétations



On peut voir qu’entre 1939 et 1945, le nombre d’écrasements d’avions a augmenté de façon significative. Des vols plus fréquents ainsi que des écrasements lors de la deuxième guerre mondial peuvent probablement expliqués cette augmentation. Par contre, On voit qu’à travers le temps, le nombre d’écrasements d’avions ne diminue pas. Cela semble plutôt stable, même si on pourrait porter à croire que les avancées technologiques réduiraient le nombre d’écrasements.

### Améliorations possibles

A rajouter….



## « Nombre de fatalités par année à travers le temps»

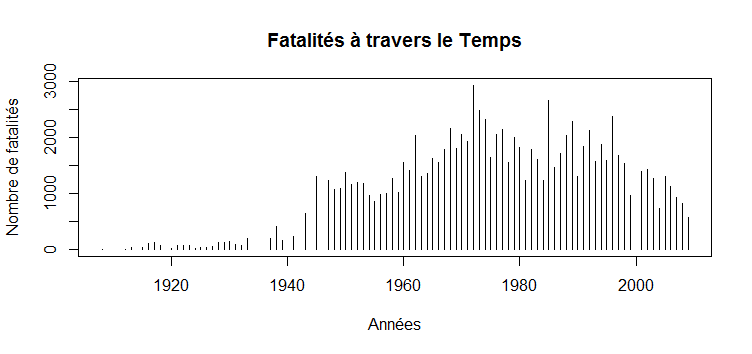
### Description

A rajouter…

### Approche de résolution

* Création d’un data.frame avec quelques colonnes du jeu de données initiales et les années
* La définition du nom des colonnes du data.frame appelé table\_de\_fatalites\_par\_annee
* Utilisation d’une boucle pour trouver la somme du nombre de fatalités pour chaque année
* L’avant-dernière étape de préparation des données implique la création d’un data.frame avec chaque valeur de somme par années et les années
* La dernière étape de préparation des données comprend la création d’un graphique à l’aide la fonction « plot »

Résultats et interprétations



Améliorations possibles

A rajouter…

## « Nombre de fatalités par location»

### Description

A rajouter…

Approche de résolution

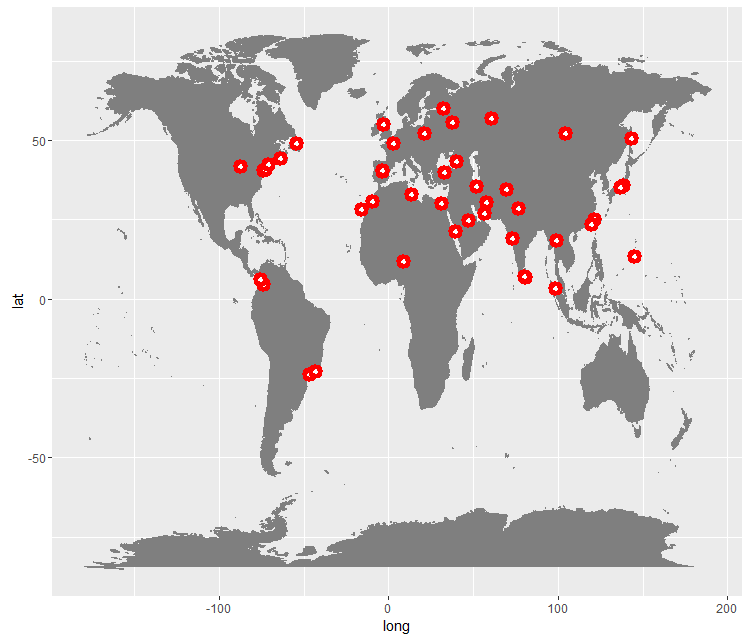
* Utilisation d’une boucle pour trouver la somme du nombre de fatalités pour chaque location
* L’organisation de la table en ordre décroissant par nombre de fatalités par location
* L’extrait du top 50 des locations avec le plus grand nombre de fatalités par location
* L’utilisation d’une boucle pour transformer en « string » chacune des instance de la variable « location » de la table top\_50\_location\_par\_fatalites
* La dernière étape de préparation des données comprend l’utilisation de la fonction « afficher\_map » pour faire afficher sur un map le top 50 des locations avec le plus grand nombre de fatalités.

Résultats et interprétations

Améliorations possibles

A rajouter…

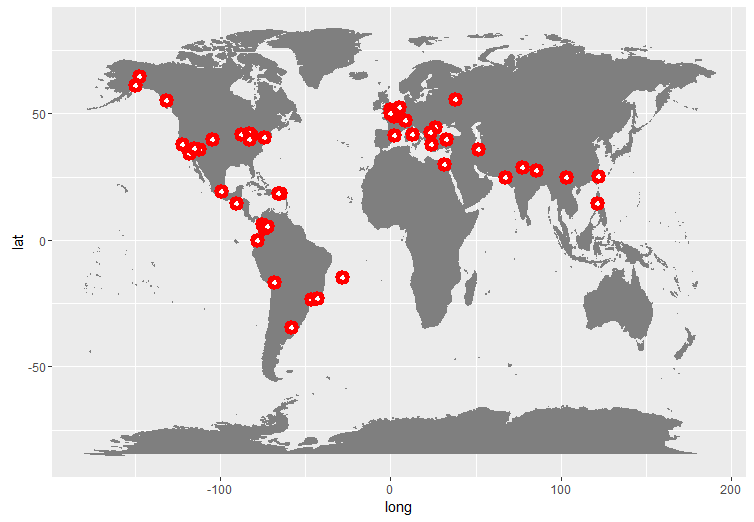
À noter que cinq valeurs qui commencent par un énoncé autre qu’un nom de ville n’affiche pas sur le map. Par exemple, les valeurs qui comment avec «Near, AtlanticOcean, off » n’ont pas retournées de coordonées géographiques.



Pour la question # , le nombre de crash par « location », la préparation des données inclut les étapes suivantes :

1. L’utilisation de la fonction trouver\_fréquence pour trouver le nombre de crashs par location
2. La création d’un data.frame pour combiner la fréquence par « location » et les «  locations »
3. L’utilisation de la fonction « order » pour mettre en ordre décroissant du nombre de crash par location
4. L’extrait du top 50 des « locations » avec la plus grande fréquence de nombres crashs
5. L’étape finale comprend l’utilisation de la fonction « afficher\_map » pour faire afficher sur un map les villes avec le plus grand nombre de crashs.

À noter que la fonction afficher\_map omet d’afficher une valeur « AtlantiOcean, 110 miles West of Ireland ». Ceci est dû au fait que la valeur entrée n’est pas un nom de ville exact, donc il est difficile de retrouver les coordonnées géographiques de cette valeur.



## « Quels sont les endroits les plus dangereux»

### Description

Le but de cette question était de regarder quels sont les endroits où un crash d’avion a le plus de chances d’arriver.

### Approche de résolution

### Résultats et interprétations

Premièrement, regardons la Figure 1 qui présente le top 10 des endroits où sont survenus des accidents d’avions. Nous voyons que les États-Unis sont premiers de loin, suivit du Brésil, de la Russie, du Canada. Il est intéressant de remarquer que la plupart des pays dans ce TOP 10 sont des pays ayant un transport aérien développé (dans le sens de plusieurs décollages et arrivés par jour) et possèdent pour la plupart un constructeur d’avion.

Un travail de suivi sera de vérifier s’il existe une corrélation entre le fait qu’un pays possède un constructeur aéronautique et le nombre de crash d’avion.

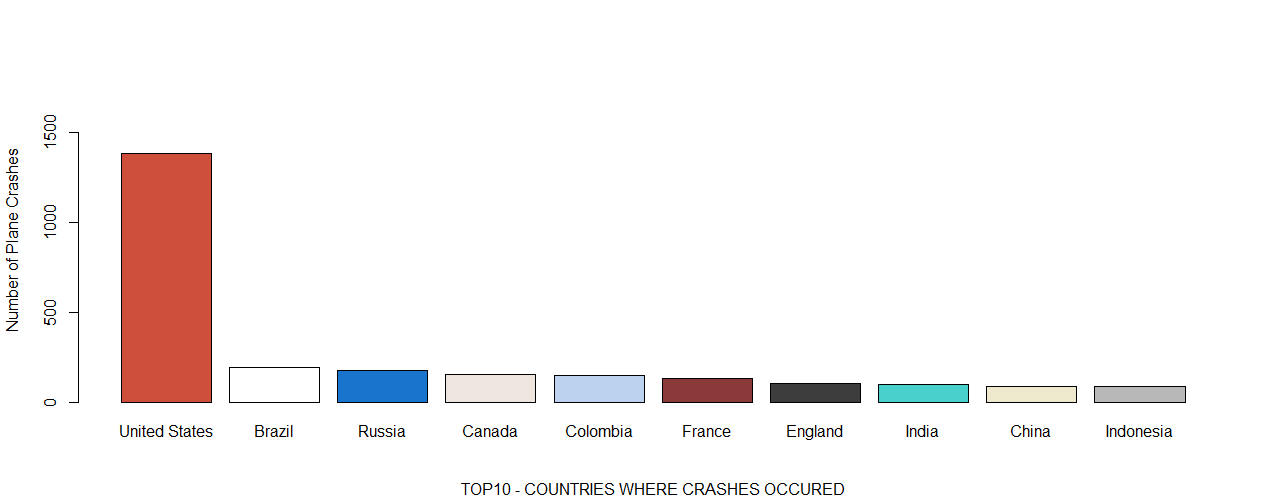


Figure 1

Dans les Figure 2 et Figure 3, le nombre d’accidents est divisé selon que l’avion était un avion de type militaire ou commercial.

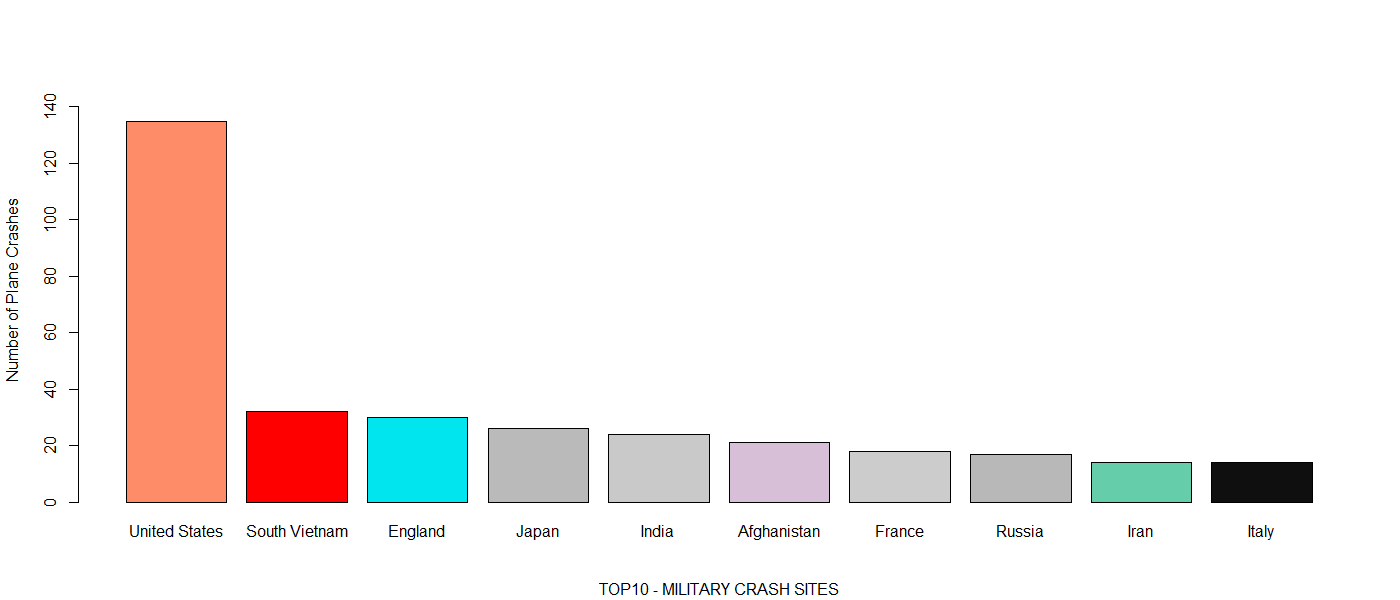
j

Figure 2

Ici, les États Unis sont toujours premiers, cependant, il est intéressant de voir que le deuxième pays pour les crashs militaires est le Sud Vietnam. Ceci nous démontre dans la collecte du jeu de données, plusieurs pays cachent l’information lorsque des crashs de nature militaires (ou même civil) surviennent. Il est important pour eux toujours présenter leur pays sous un bon jour. D’après les auteurs, le Sud Vietnam occupe cette position du fait que les écrasements d’avions ont été répertoriés durant la guerre du Vietnam, ce qui n’est pas le cas pour plusieurs autres conflits.

Un des auteurs avait espéré que la Figure 3 démontrerait l’existence du fameux triangle des Bermudes <https://en.wikipedia.org/wiki/Bermuda_Triangle> . Ceci n’est pas le cas, les données compilées n’ont aucun pays de ce triangle parmi les endroits où surviennent le plus d’accidents d’avion.

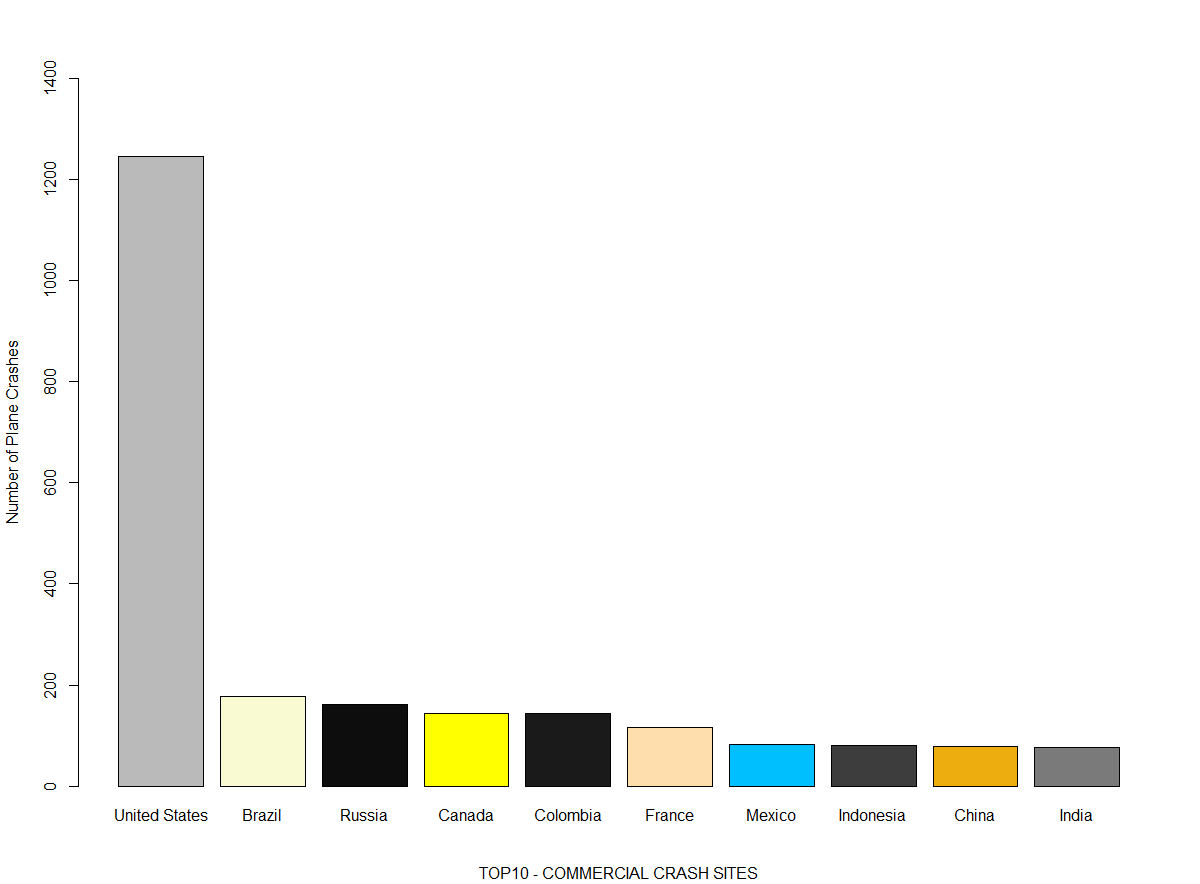
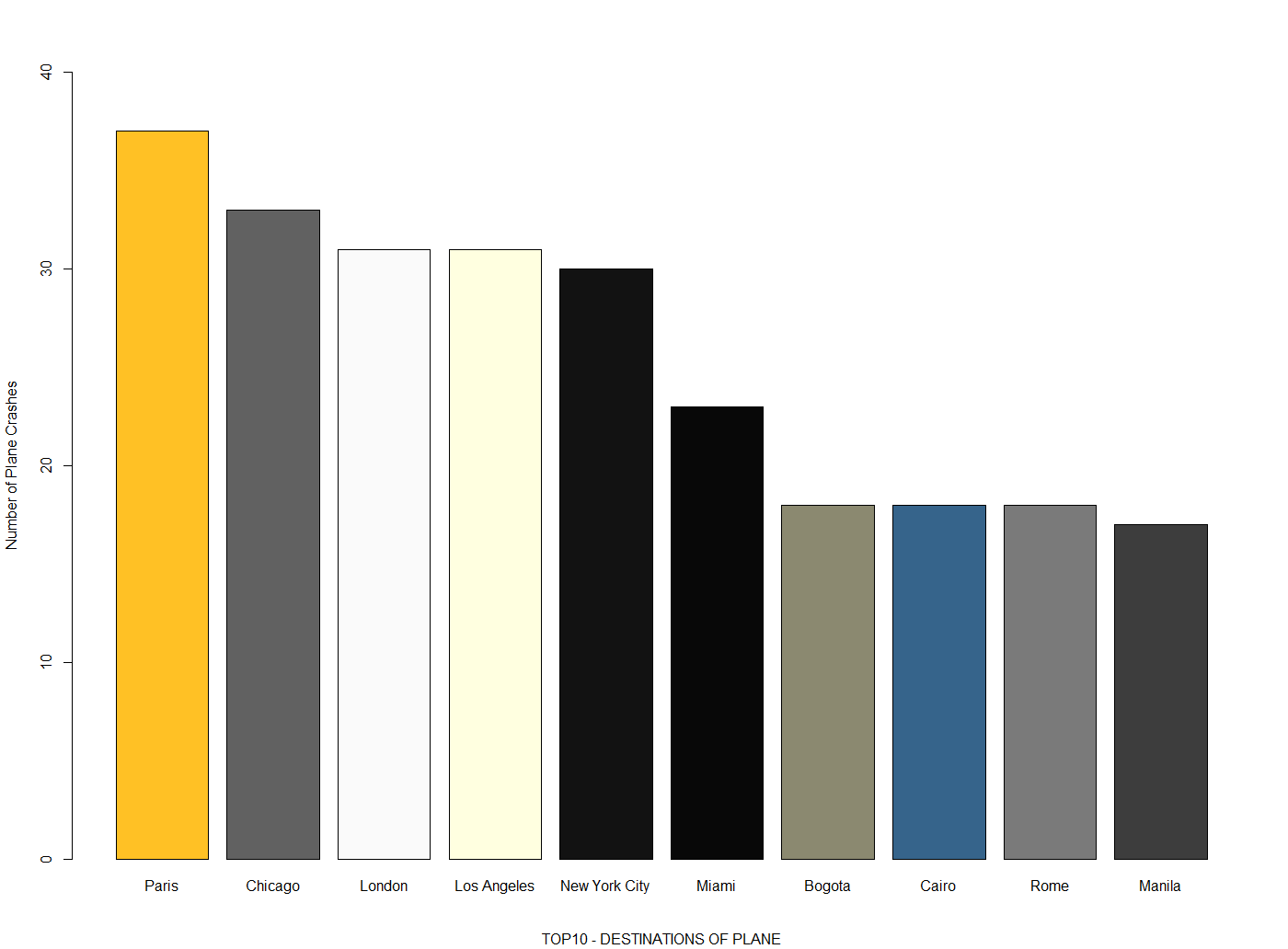
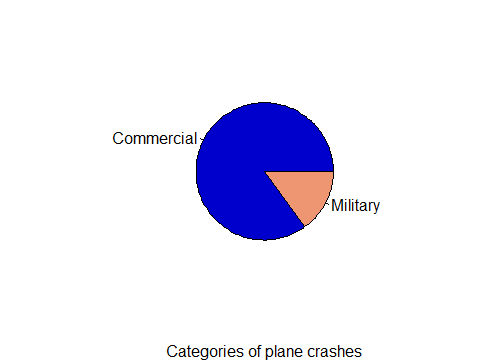


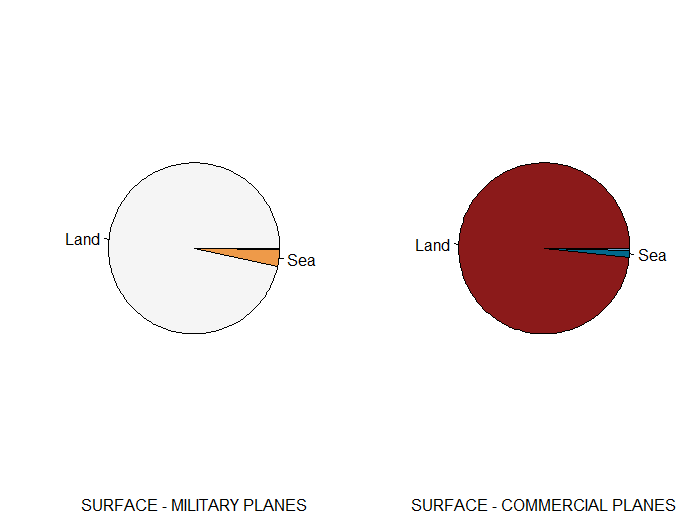
Figure 3

La figure suivante nous montre que la plupart des accidents surviennent lors des vols reliant de grandes villes de ce monde. En effet, le sens commun nous dit que plus il y a de vols vers une destination, plus il y aura des accidents parmi les vols vers cette destination. Un exercice statistique serait de démontrer cette inférence.

## Risque selon les types d’avion (william.R):



Le jeu de données nous montre que la plupart des avions à être impliquer dans des écrasements sont des avions commerciaux. Il est important de mentionner qu’il existe un biais. En effet, comme mentionné plus haut dans le rapport, la plupart des pays ne signale pas lorsqu’un de leur avion militaire s’écrase.



Les accidents d’avions surviennent dans la grande majorité des cas sur la terre ferme, contrairement aux présuppositions d’un des auteurs du rapport.

## C:\Users\tankou\Documents\GitHub\hecLogicielStatistiques\Rplot07.png

## Chance de survie (william.R):

Le dernier graphique nous montre l’estimation de survie lors d’un crash selon que l’avion soit militaire ou commercial et que le crash survient lors d’un entrainement ou non.

## « catégories des raisons des crash»

### Description

Le but de cette question est d’avoir une idée globale des raisons pour lesquelles les crash arrivent.

### Approche de résolution

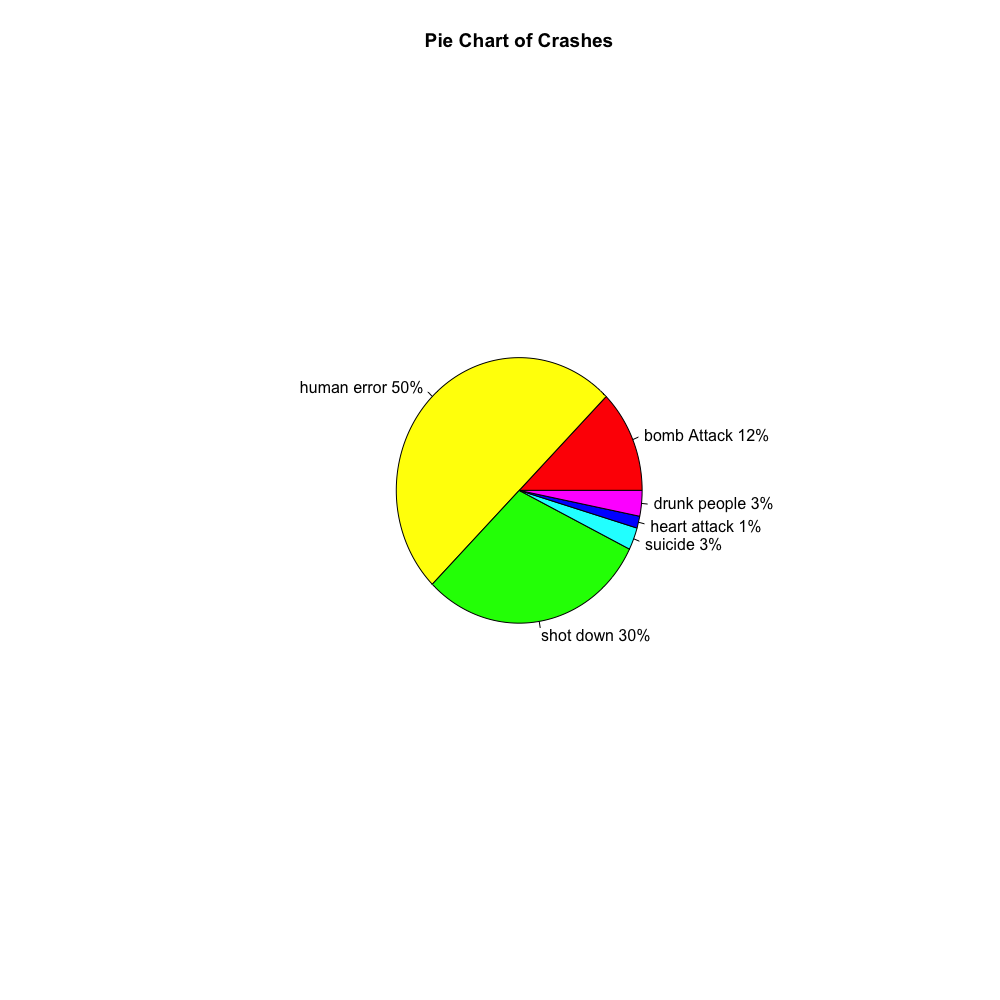
Pour pouvoir avoir une idée globale des raisons, voici l’approche adoptée :

* Création des taxonomies pour les catégories :

BOMB\_ATTACK\_CATEGORIES, ERROR\_CATEGORIES, SHOT\_DOWN\_CATEGORIES, SHOT\_DOWN\_CATEGORIES, SHOT\_DOWN\_CATEGORIES DRUNK\_ATTACK\_CATEGORIES

* Parcours du data frame et Identification de catégorie pour chaque entrée
* L’approche choisie est une approche naïve qui consiste à trouver une série de mots clés dans la description du crash.

### Résultats et interprétations



### Améliorations possibles

L’approche de resolution choisie n’est pas optimale car elle peut être erronnée. C’est une approche naïve. L’idéal serait d’aller vers une approche de “Text analysis”. Cette technique donnera beaucoup plus de résultats et serait beaucoup plus efficace.

## « Quelle est la période la plus dangereuse (Janvier-Juin) et quelles sont ces destinations»

### Description

Le but est de savoir la période de l’année la plus dangereuse ainsi que les destinations concernées.

### Approche de résolution

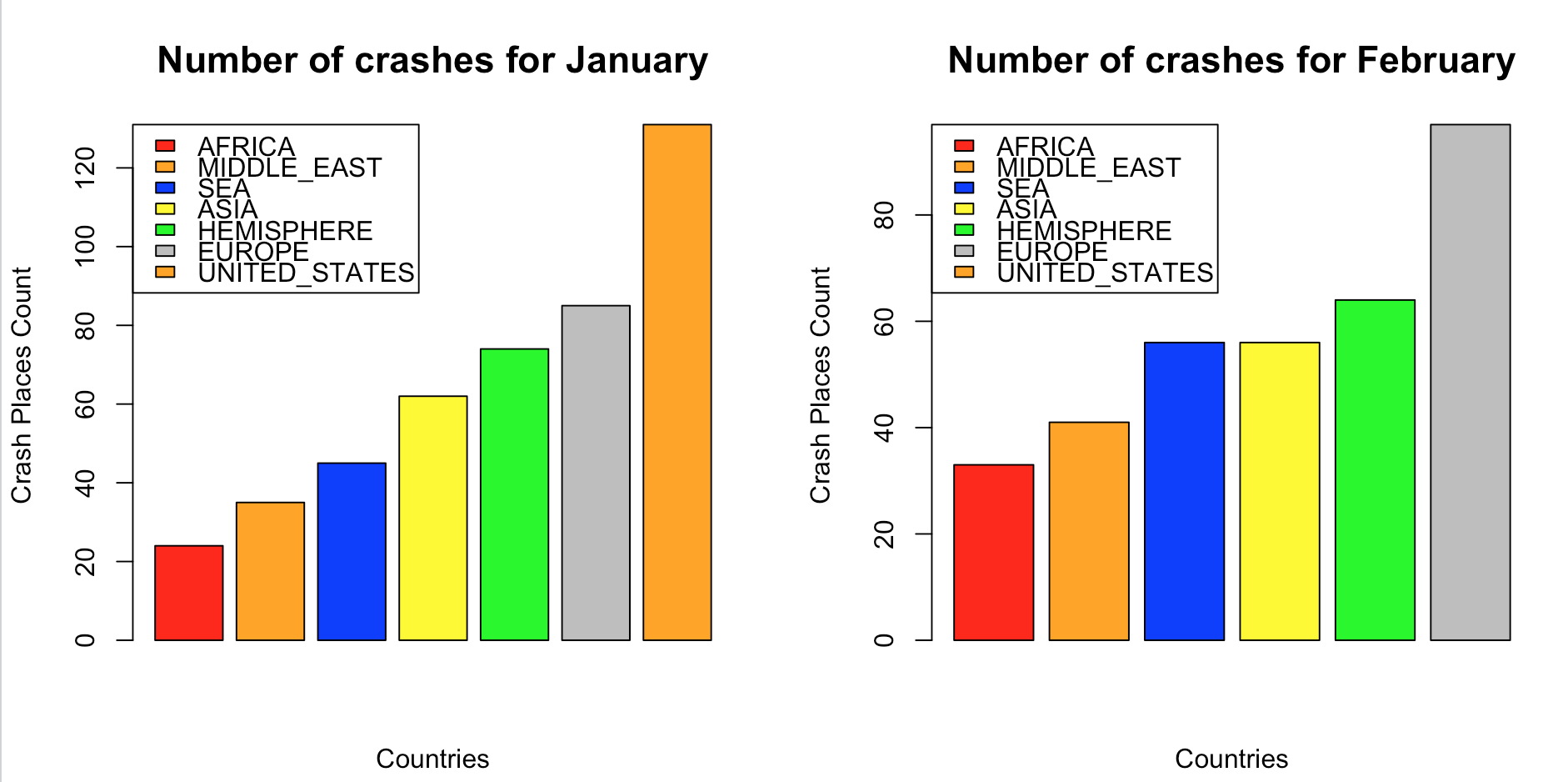
Elle se détaille comme suit:

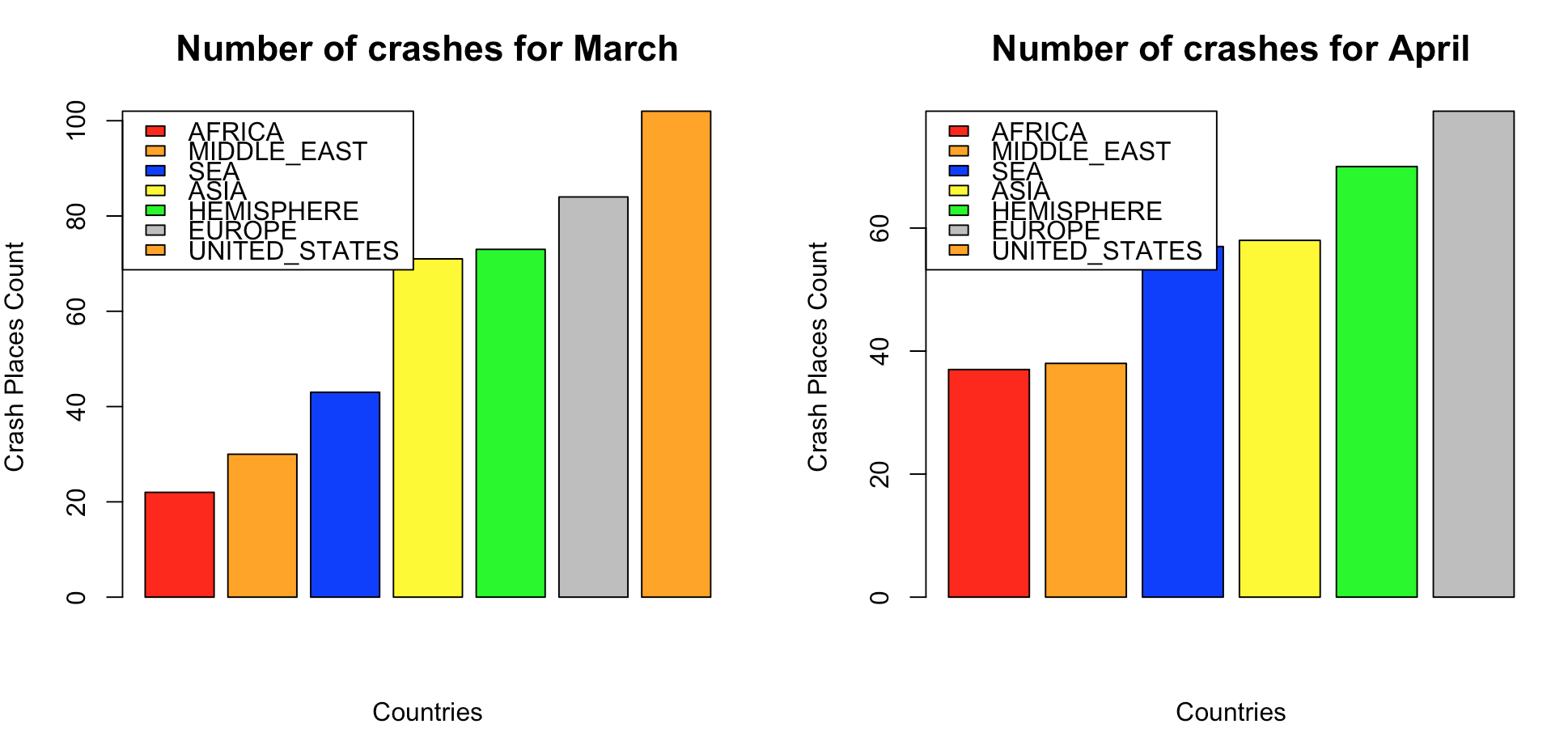
* Correction des noms de pays erronés
* Élimination des entrées vides ou manquantes
* Création des taxonomies pour les locations :

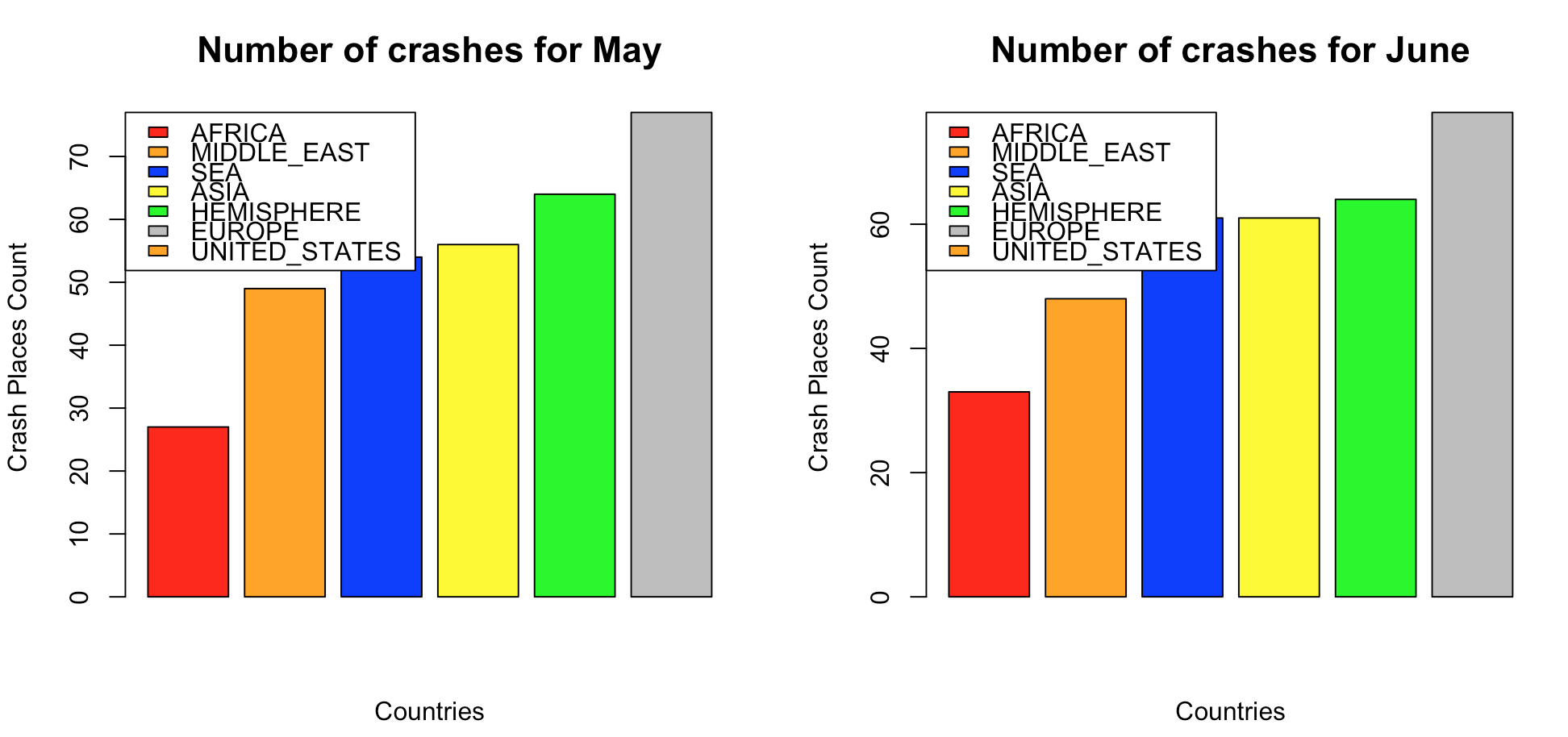
CANADA, AFRICAN\_COUNTRIES, WESTERN\_HEMISPHERE, SEA\_WORLD, EUROPE\_COUNTRIES EUROPE\_COUNTRIES, SOUTHEAST\_ASIA\_COUNTRIES, US\_CITIES, MIDDLE\_EAST

* Parcours du data frame et Identification de location pour chaque entrée
* L’approche choisie est une approche naïve qui consiste à trouver une série de mots clés dans la colonne location de l’entrée.

### Résultats et interprétation







### Améliorations possibles

## « Quels sont les vols qui ont fait plus de morts»

### Description

Le but de la question est de savoir au cours du temps, les opérateurs qui ont fait plus de morts.

### Approche de résolution

Elle se détaille comme suit:

* Élimination des entrées vides ou manquantes
* Création des taxonomies pour les opérateurs:

AIR\_FRANCE, AIR\_CANADA, TRANS\_CANADA, AIR\_ONTARIO, AIR\_INDIA, AIR\_CARAIBES, AIR\_MADAGASCAR, AIR\_NIAGARA, AIR\_GUADELOUPE, AIR\_AMERICA, AIR\_MALI,

DEUTSCHE\_LUFTHANSA, CHINA\_AIRLINES

* Parcours du data frame et Identification de l’opérateur pour chaque entrée
* L’approche choisie est une approche naïve qui consiste à trouver une série de mots clés dans la colonne Operateur de l’entrée.

### Résultats et interprétation

### 

### Améliorations possibles

* L’analyse est naïve dans la recherche de l’opérateur. Ceci peut être une source d’erreur dans l’analyse.
* Le type d’avion devrait aussi rentrer dans l’équation car plus un opérateur a de gros avions plus il a de chances d’avoir des bilans lourds quant au nombre de victimes.