PREDICTING COMPLIANCE: A PEOPLE ANALYTICS APPROACH TO PROACTIVE INTERVENTION

OLAPEJU ESUOLA, MSBA

APRIL 15, 2025

Abstract

This project explores the intersection of people analytics and quality compliance, focusing on how employee behavior impacts the success of quality initiatives. The key challenge addressed is ensuring consistent compliance with quality procedures, particularly during the implementation of process improvements. By analyzing employee performance indicators such as sales targets, customer satisfaction, and working patterns, this study identifies behavioral and operational factors influencing compliance outcomes. A dataset containing policy compliance records and performance metrics is used to uncover patterns and predictors of non-compliance. The findings emphasize that effective compliance begins with people and that leveraging data can support proactive, people-centered quality assurance strategies.

Introduction

People analytics is a rapidly growing field that applies data-driven insights to improve workforce management and organizational performance. In the context of quality management, it offers unique opportunities to enhance compliance with procedural standards. As organizations continuously evolve, the introduction of new quality processes becomes inevitable—yet one of the most significant challenges is ensuring employee adherence, particularly in the face of resistance to change. The alignment between HR and quality departments plays a critical role in overcoming these challenges and embedding compliance behaviors into workplace culture (Tariq, 2024).

This project investigates the use of people analytics to support HR and quality department collaboration in improving employee compliance. The dataset under review offers insights into employee performance, engagement, and policy adherence—factors critical to evaluating and supporting the successful implementation of new quality procedures (Buchanan & Kittie, 2023).

Literature Review

The intersection of people analytics and quality compliance has garnered increased academic and industry attention. Huselid (2023) emphasizes the strategic value of workforce data in shaping performance management systems that drive operational outcomes. His work supports the notion that predictive models of employee behavior can reduce quality failures and elevate organizational agility.

Buchanan and Kittie (2023) explore organizational behavior frameworks that link employee engagement, training, and role clarity to compliance outcomes. Their findings reinforce that cultural and structural enablers—such as managerial communication and feedback loops—play a significant role in embedding new processes successfully.

Tariq (2024) highlights the critical role of HR in managing change and ensuring employees align with quality expectations during system transitions. The study identifies frontline staff as both the implementers and enablers of quality, suggesting that their behavior should be continuously monitored and supported using data-driven approaches.

Together, these studies provide a strong foundation for this project’s focus: using people analytics to understand and improve employee compliance during periods of procedural change.

Problem Statement

Ensuring compliance with quality procedures remains a core challenge for organizations, especially during the rollout of new processes or continuous improvement initiatives. Resistance to change, lack of clarity in expectations, and insufficient engagement among frontline staff often hinder full adherence. From a quality specialist's perspective, it is clear that successful compliance requires more than documented SOPs; it demands behavioral alignment supported by HR intervention.

Despite the potential of people analytics to provide actionable insights into workforce behavior, many organizations still struggle to apply these tools effectively to influence compliance. The challenge intensifies when collaboration between HR and quality departments is limited, leaving gaps in training, communication, and accountability (Tariq, 2024; Huselid, 2023).

Proposed Solution

To improve compliance outcomes during process changes, this project proposes a data-informed collaboration between HR and quality departments. People analytics can help identify patterns in employee behavior—such as low engagement, missed targets, or poor customer satisfaction—that signal a risk of non-compliance.

By leveraging these insights, targeted interventions such as coaching, retraining, or communication reinforcement can be implemented. This approach supports a more adaptive, proactive compliance culture. Through better alignment of HR strategies and quality objectives, organizations can promote buy-in, reduce resistance to change, and enhance overall quality performance (Buchanan & Kittie, 2023; Kittie, 2023).

Methodology

This project uses a dataset titled Policy_Compliance_Dataset_Updated.csv, which includes variables related to employee performance, task execution, and policy adherence. The dataset captures metrics such as working days, sales targets, actual performance, customer satisfaction scores, and reasons for non-compliance.

By analyzing these variables, patterns in compliance behavior can be identified. The analysis focuses on understanding how factors like underperformance, customer dissatisfaction, and absenteeism correlate with policy non-compliance. The results will inform strategies for improving employee accountability and compliance during process transitions.

Research Design

This study adopts a quantitative exploratory design using secondary data to assess employee compliance behaviors in relation to performance and engagement metrics. The research follows a structured process beginning with data cleaning, exploratory analysis, and transformation, followed by model building and interpretation. The aim is to identify patterns and predictive signals that influence compliance with quality procedures.

Data Source

The dataset used in this study, titled Employee Policy Compliance Dataset, was sourced from Kaggle. It contains structured, publicly available data that simulates employee performance, policy compliance, and behavioral indicators in a workplace setting.

The dataset includes the following key variables:

- Employee_ID and Name – Unique identifiers for individual employees
- Working_Days – Number of days each employee worked per month
- Target_Sales and Actual_Sales – Metrics used to assess individual productivity
- Customer_Satisfaction_Score – Numeric feedback representing service or interaction quality
- Policy_Compliance – Indicates whether the employee adhered to organizational policies
- Low_Working_Days, Target_Not_Met, and Low_Customer_Satisfaction – Binary flags for underperformance
- Non_Compliance_Reason – Categorical field explaining the reason for non-compliance
- Month – Temporal variable allowing for trend and time-series analysis

This dataset is suitable for the study’s objective of exploring the relationship between employee performance and compliance behavior, as it provides a range of attributes relevant to both operational outcomes and policy adherence. The data structure supports the application of people analytics and predictive modeling techniques to uncover insights into quality compliance patterns in the workplace.

Data Preparation & Cleaning

To ensure accurate analysis, the dataset underwent a structured preparation process. This included:

- Data inspection to understand variable types and detect inconsistencies
- Drop rows containing missing values
- Standardization of variable formats (e.g., categorical factors, column naming)
- Validation by previewing the cleaned dataset

```
# Load required libraries
library(readr)
library(dplyr)
library(tidyr)
library(stringr)
library(ggplot2)
# Load dataset
Policy_Compliance_Dataset_Updated <- read.csv("Policy_Compliance_Dataset_Updated.csv")
# Summary stats
summary(Policy_Compliance_Dataset_Updated)
```

```
##   Employee_ID       Name            Working_Days    Target_Sales
##  Min.   :   1   Length:4000        Min.   :15.00   Min.   : 5002
##  1st Qu.:1001   Class :character   1st Qu.:18.00   1st Qu.: 8707
##  Median :2000   Mode  :character   Median :22.00   Median :12425
##  Mean   :2000                      Mean   :22.41   Mean   :12497
##  3rd Qu.:3000                      3rd Qu.:26.00   3rd Qu.:16368
##  Max.   :4000                      Max.   :30.00   Max.   :19996
##   Actual_Sales   Customer_Satisfaction_Score Policy_Compliance
##  Min.   : 4001   Min.   :2.000               Length:4000
##  1st Qu.: 8420   1st Qu.:2.800               Class :character
##  Median :12998   Median :3.500               Mode  :character
##  Mean   :12943   Mean   :3.494
##  3rd Qu.:17453   3rd Qu.:4.200
##  Max.   :22000   Max.   :5.000
##  Low_Working_Days   Target_Not_Met     Low_Customer_Satisfaction
##  Length:4000        Length:4000        Length:4000
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Non_Compliance_Reason     Month
##  Length:4000          Length:4000
##  Class :character     Class :character
##  Mode  :character     Mode  :character
##
##
##
```

The summary statistics revealed a balanced dataset with consistent numeric ranges and no extreme outliers across key performance variables. Categorical and logical variables were well-distributed, with compliance-related flags providing clear segmentation for further analysis.

```
# Rename dataset
df <- Policy_Compliance_Dataset_Updated

df$Non_Compliance_Reason[df$Non_Compliance_Reason == "" & df$Policy_Compliance == "Yes"] <- "Compliant"


# Check total missing values per column
colSums(is.na(df))
```

```
##                Employee_ID                       Name
##                          0                          0
##               Working_Days               Target_Sales
##                          0                          0
##               Actual_Sales Customer_Satisfaction_Score
##                          0                          0
##          Policy_Compliance           Low_Working_Days
##                          0                          0
##             Target_Not_Met   Low_Customer_Satisfaction
##                          0                          0
##      Non_Compliance_Reason                      Month
##                          0                          0
```

During the cleaning process, it was discovered that 640 rows had missing values in the Non_Compliance_Reason column. However, further inspection revealed that these entries corresponded to employees who were fully compliant, and thus no reason for non-compliance was applicable. These missing values were replaced with the label “Compliant” to preserve the completeness of the dataset without removing valid records. A follow-up check confirmed that the updated dataset contains zero missing values across all columns.

```
# Rename columns to remove spaces and apply consistent formatting
names(df) <- gsub(" ", "_", names(df))

# Convert character columns to factors where appropriate
df$Policy_Compliance <- as.factor(df$Policy_Compliance)
df$Month <- as.factor(df$Month)
df$Name <- as.factor(df$Name)
df$Non_Compliance_Reason <- as.factor(df$Non_Compliance_Reason)

# Ensure binary fields are formatted as 0/1 instead of logical
df$Low_Working_Days <- ifelse(df$Low_Working_Days == "True", 1, 0)
df$Target_Not_Met <- ifelse(df$Target_Not_Met == "True", 1, 0)
df$Low_Customer_Satisfaction <- ifelse(df$Low_Customer_Satisfaction == "True", 1, 0)
```

At this stage, data standardization was performed to ensure consistency in data types and formatting. Numerical normalization was deferred until the feature engineering phase to allow for tailored scaling of model-specific variables.

```
# Preview cleaned dataset
head(df, 5)
```

```
##   Employee_ID        Name Working_Days Target_Sales Actual_Sales
## 1           1 Ahmed Tariq           21        14435        19470
## 2           2 Usman Iqbal           18         9998         9968
## 3           3 Fatima Khan           27         5162        11493
## 4           4  Zain Tariq           29         6974         8103
## 5           5 Ahmed Malik           25         8291         6076
##   Customer_Satisfaction_Score Policy_Compliance Low_Working_Days Target_Not_Met
## 1                         2.3                No                0              0
## 2                         3.0                No                1              0
## 3                         3.3                No                0              0
## 4                         2.8                No                0              0
## 5                         3.7                No                0              1
##   Low_Customer_Satisfaction                        Non_Compliance_Reason
## 1                         1                    Low Customer Satisfaction
## 2                         1 Low Working Days, Low Customer Satisfaction
## 3                         1                    Low Customer Satisfaction
## 4                         1                    Low Customer Satisfaction
## 5                         1   Target Not Met, Low Customer Satisfaction
##      Month
## 1    March
## 2 December
## 3 February
## 4      May
## 5 February
```

Following a structured data preparation process—including the handling of context-based missing values, standardization, and formatting—a clean and analysis-ready dataset was obtained. Missing entries in the Non_Compliance_Reason column were not removed, as they represented employees who were fully compliant; these were relabeled as “Compliant” to preserve valid observations. Variables were converted to their appropriate types, with categorical fields set as factors and binary indicators standardized to 0 and 1. This finalized dataset provides a reliable foundation for exploratory analysis, feature engineering, and predictive modeling aimed at understanding employee compliance behavior.

Data Transformation

To enhance the depth of exploratory analysis, several data transformations were performed. A Performance_Gap variable was introduced to assess how far employees deviated from their sales targets. A binary Compliance_Status flag was created for simplified analysis, and satisfaction scores were grouped into categories to support comparison. These transformations allow for more targeted exploration of behavioral patterns influencing compliance.

```
# Create a performance gap column
df <- df %>%
  mutate(Performance_Gap = Actual_Sales - Target_Sales)

# Create a compliance status flag column
df <- df %>%
  mutate(Compliance_Status = ifelse(Policy_Compliance == "Yes", 1, 0))

# Categorize customer satisfaction
df <- df %>%
  mutate(Satisfaction_Level = case_when(
    Customer_Satisfaction_Score < 3 ~ "Low",
    Customer_Satisfaction_Score >= 3 & Customer_Satisfaction_Score < 4 ~ "Moderate",
    Customer_Satisfaction_Score >= 4 ~ "High"
  ))
```

```
# Set your employee-themed color palette
my_colors <- c(
  "Yes" = "#2C7BB6",      # Blue for compliant
  "No" = "#D7191C",       # Red for non-compliant
  "Neutral" = "#FDAE61",  # Amber
  "Highlight" = "#ABDDA4" # Light green
)

# Global plot theme
theme_set(
  theme_minimal(base_size = 12) +
    theme(
      plot.title = element_text(face = "bold", color = "#333333"),
      axis.title = element_text(face = "bold"),
      legend.title = element_text(face = "bold"),
      strip.text = element_text(face = "bold", color = "#2C7BB6")
    )
)
```

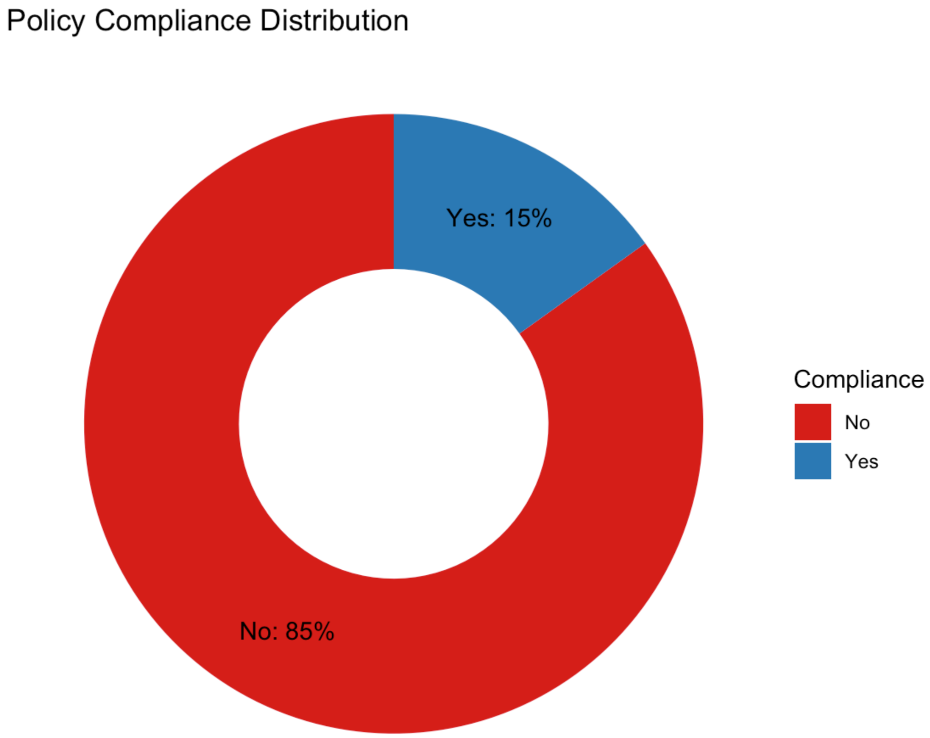
A people-centered visual theme was applied throughout the project to reflect the employee-focused nature of the analysis. Soft blues and warm reds were used to distinguish compliant vs. non-compliant behavior, while neutral tones helped highlight flags and moderate outcomes. This approach supports clear storytelling around workplace behavior and policy adherence.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed to uncover trends, relationships, and distribution patterns within the dataset. This step provides essential context for understanding the factors associated with employee compliance behavior, including performance metrics, customer satisfaction levels, and monthly variations.

Visual 1: Donut Chart – Compliance Proportion

```
df %>%
  count(Policy_Compliance) %>%
  mutate(prop = n / sum(n),
         label = paste0(Policy_Compliance, ": ", round(prop * 100), "%")) %>%
  ggplot(aes(x = 2, y = prop, fill = Policy_Compliance)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  scale_fill_manual(values = my_colors) +
  geom_text(aes(label = label), position = position_stack(vjust = 0.5)) +
  xlim(0.5, 2.5) +
  theme_void() +
  labs(title = "Policy Compliance Distribution",
       fill = "Compliance")
```

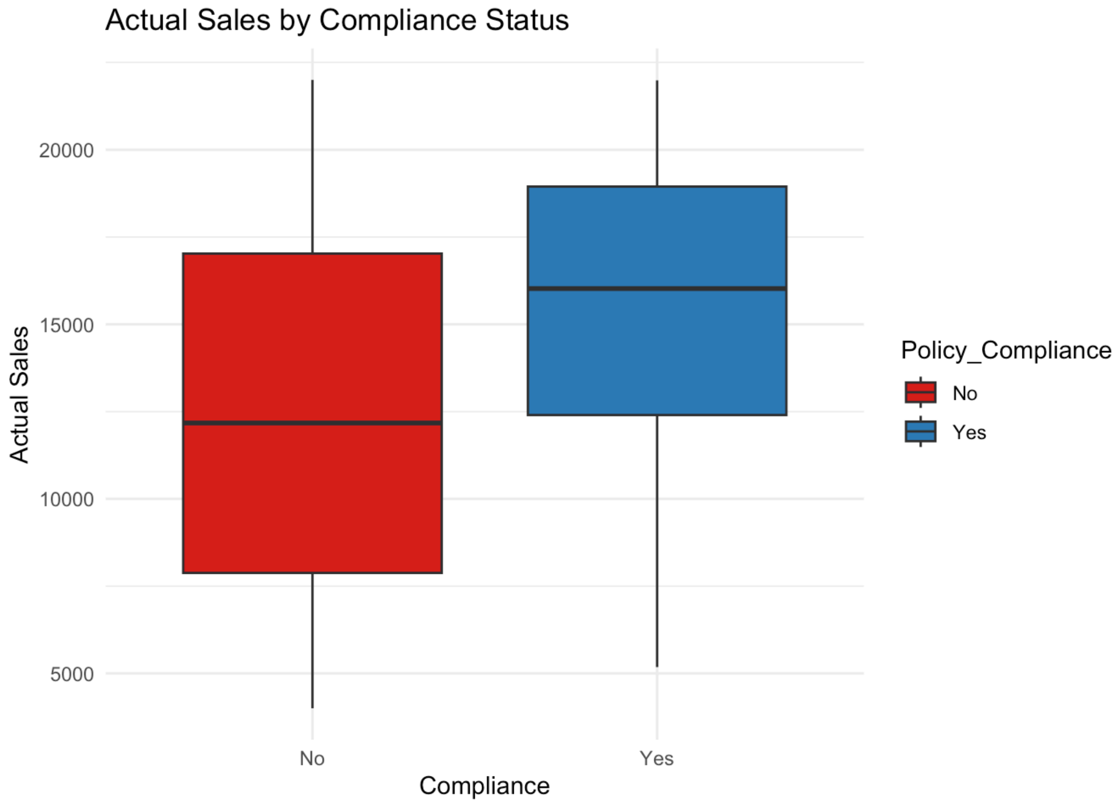


Result: The donut chart shows that out of 4,000 employees, 85% were non-compliant and only 15% were compliant.

Interpretation: This stark imbalance suggests a potential gap in training, communication, or engagement. The majority of the workforce is not meeting expected policy standards, warranting proactive strategies for compliance improvement.

Visual 2: Box Plot – Actual Sales by Compliance Status

```
# Actual Sales by Compliance Status
ggplot(df, aes(x = Policy_Compliance, y = Actual_Sales, fill = Policy_Compliance)) +
  geom_boxplot() +
    scale_fill_manual(values = my_colors) +
  labs(title = "Actual Sales by Compliance Status",
       x = "Compliance",
       y = "Actual Sales") +
  theme_minimal()
```

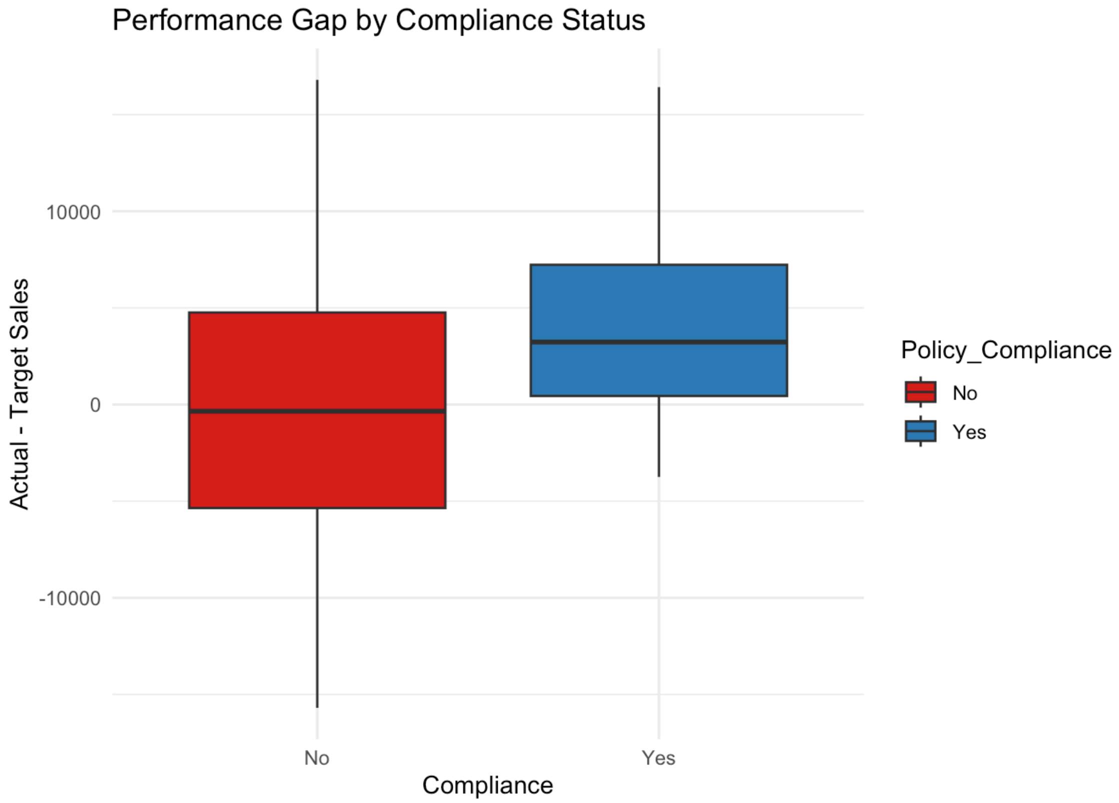


Result: Compliant employees had a higher median actual sales (~16,000) compared to non-compliant employees (~12,000).

Interpretation: Employees who comply with policies also tend to perform better in sales, suggesting a positive relationship between compliance and productivity. This may reflect better discipline, process adherence, or engagement.

Visual 3: Box Plot – Performance Gap by Compliance

```
ggplot(df, aes(x = Policy_Compliance, y = Performance_Gap, fill = Policy_Compliance)) +
  geom_boxplot() +
  scale_fill_manual(values = my_colors) +
  labs(title = "Performance Gap by Compliance Status",
       x = "Compliance", y = "Actual - Target Sales") +
  theme_minimal()
```

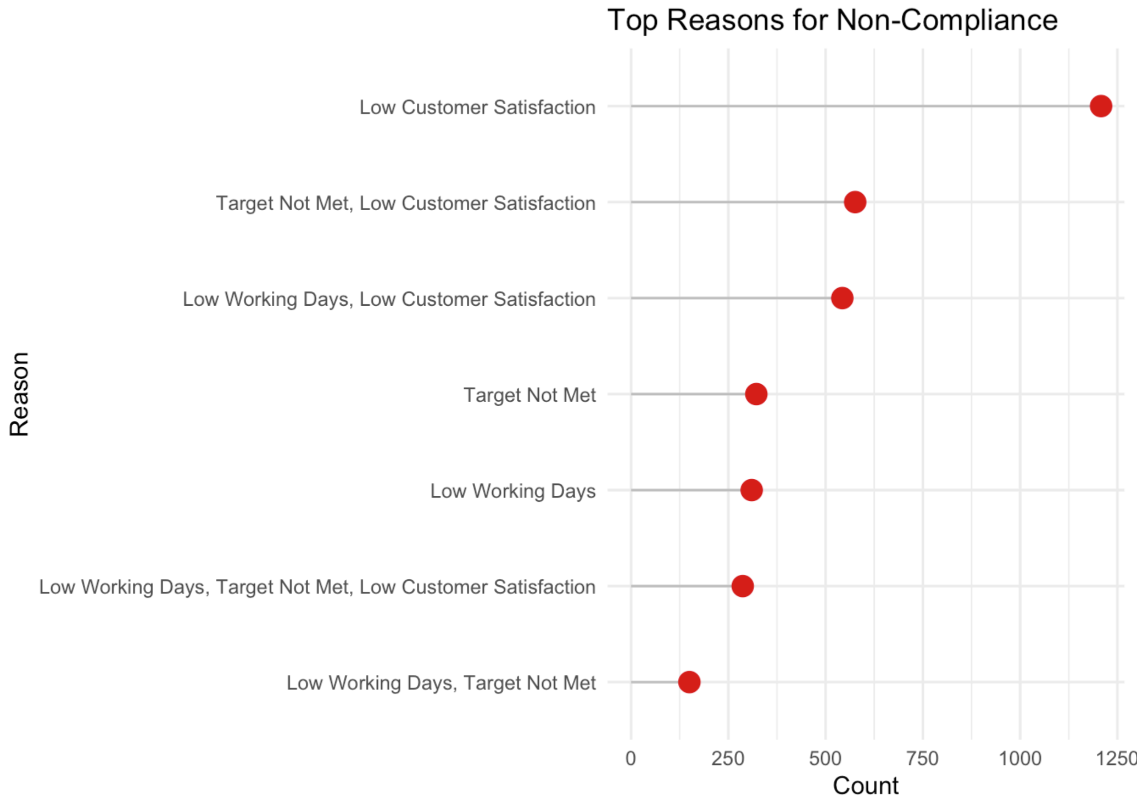


Result: The performance gap (Actual - Target Sales) for compliant employees was generally positive, while non-compliant employees had negative gaps.

Interpretation: Non-compliant employees are more likely to underperform relative to their targets. This reinforces the importance of monitoring early signs of underperformance as predictors of future non-compliance.

Visual 4: Lollipop Chart – Top 10 Reasons for Non-Compliance

```
df %>%
  filter(Policy_Compliance == "No") %>%
  count(Non_Compliance_Reason, sort = TRUE) %>%
  slice_head(n = 10) %>%
  ggplot(aes(x = reorder(Non_Compliance_Reason, n), y = n)) +
  geom_segment(aes(xend = Non_Compliance_Reason, yend = 0), color = "gray") +
  geom_point(color = "#D7191C", size = 4) +
  coord_flip() +
  labs(title = "Top Reasons for Non-Compliance",
       x = "Reason", y = "Count") +
  theme_minimal()
```



Result: The most common reasons include Low Customer Satisfaction, Target Not Met, and Low Working Days, either individually or in combination.

Interpretation: This highlights that employee behavior and feedback — especially from customers — are critical indicators of compliance risk. Organizations should track these metrics continuously and treat them as early warning signs.

Visual 5 & 6: Sankey Diagram – Contributions to Compliance/Non-Compliance

```
# Count how many non-compliant employees fall into each issue category
target_not_met_noncompliant <- nrow(df[df$Policy_Compliance == "No" & df$Target_Not_Met == 1, ])
low_satisfaction_noncompliant <- nrow(df[df$Policy_Compliance == "No" & df$Low_Customer_Satisfaction ==
1, ])
low_working_days_noncompliant <- nrow(df[df$Policy_Compliance == "No" & df$Low_Working_Days == 1, ])
```

```
# Count how many compliant employees had positive attributes
target_met_compliant <- nrow(df[df$Policy_Compliance == "Yes" & df$Target_Not_Met == 0, ])
high_satisfaction_compliant <- nrow(df[df$Policy_Compliance == "Yes" & df$Low_Customer_Satisfaction == 0,
])
high_working_days_compliant <- nrow(df[df$Policy_Compliance == "Yes" & df$Low_Working_Days == 0, ])
```

```
target_met_compliant
```

```
## [1] 604
```

```
high_satisfaction_compliant
```

```
## [1] 604
```

```
high_working_days_compliant
```

```
## [1] 604
```

```
target_not_met_noncompliant
```

```
## [1] 1335
```

```
low_satisfaction_noncompliant
```

```
## [1] 2614
```

```
low_working_days_noncompliant
```

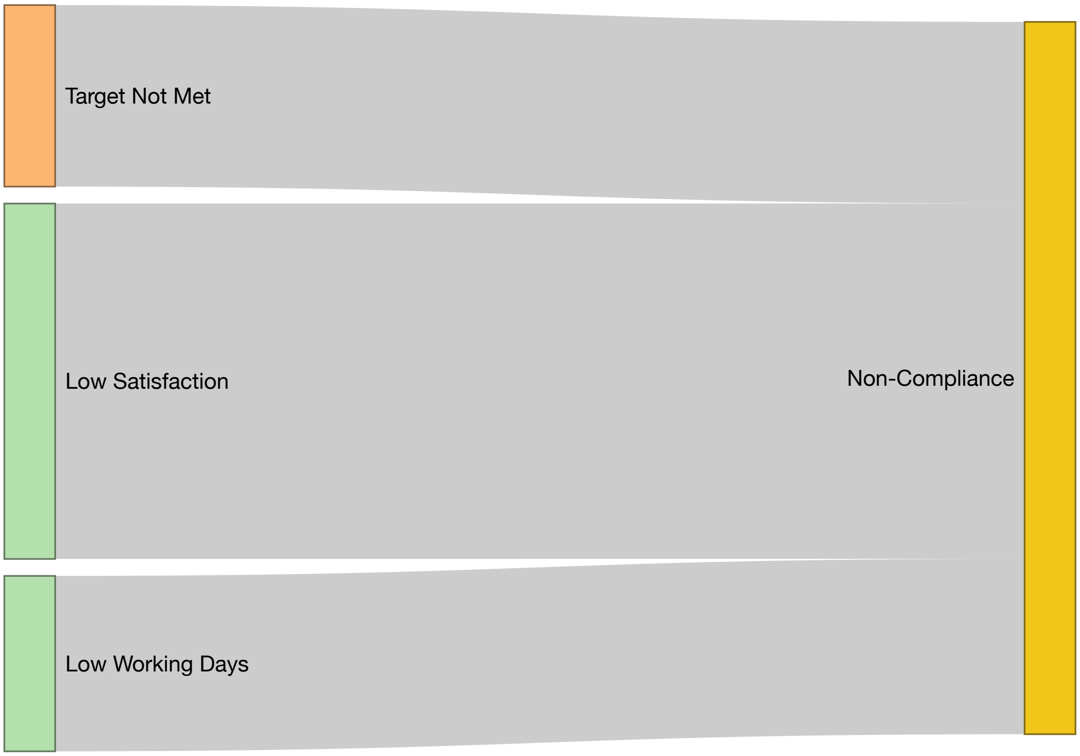
```
## [1] 1290
```

```
library(networkD3)
library(dplyr)

# Step 1: Create nodes
nodes <- data.frame(name = c("Target Not Met", "Low Satisfaction", "Low Working Days", "Non-Compliance"))

# Step 2: Create links
links <- data.frame(
  source = c(0, 1, 2),
  target = c(3, 3, 3),
  value = c(1335, 2614, 1290)
)

# Step 3: Plot Sankey
sankeyNetwork(Links = links, Nodes = nodes, Source = "source",
              Target = "target", Value = "value", NodeID = "name",
              fontSize = 13, nodeWidth = 30, colourScale = JS("d3.scaleOrdinal().range(['#FDAE61', '#ABDD
A4', '#EFC000', '#D7191C'])"))
```



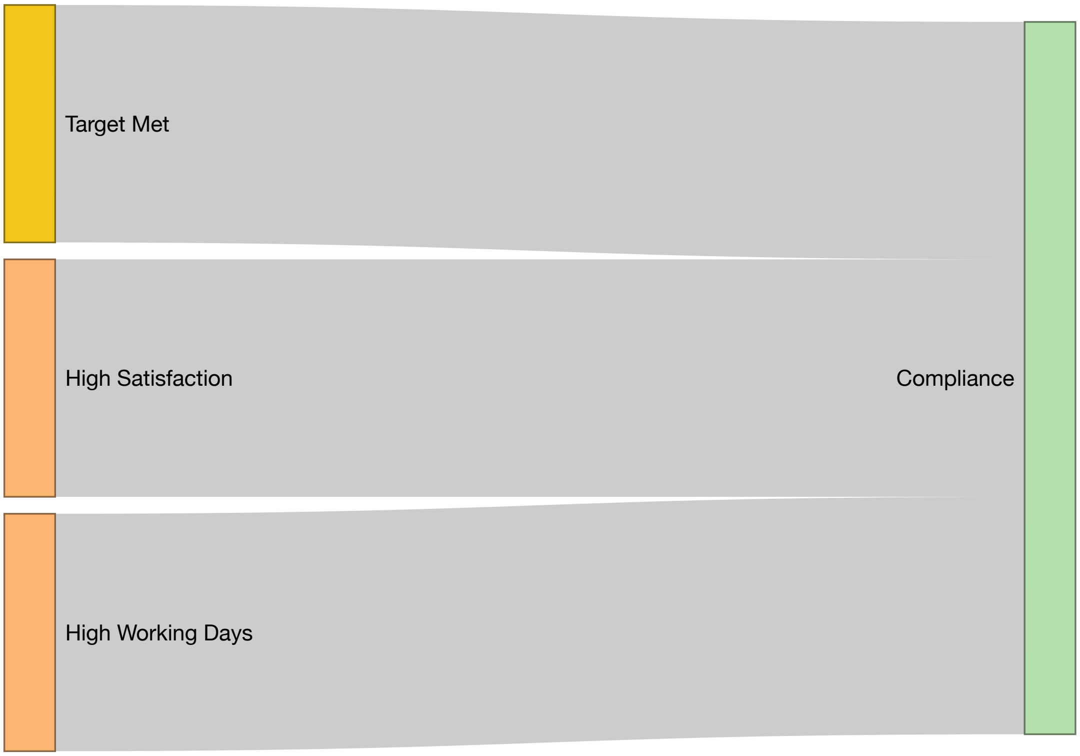
```r
# Load libraries
library(networkD3)

# Define nodes
nodes <- data.frame(name = c(
  "Target Met",
  "High Satisfaction",
  "High Working Days",
  "Compliance"
))

# Define links: source = index of node in 'nodes$name', target = index of target node
links <- data.frame(
  source = c(0, 1, 2),      # "Target Met", "High Satisfaction", "High Working Days"
  target = c(3, 3, 3),      # All point to "Compliance"
  value = c(604, 604, 604)  # Replace with your actual values if available
)

# Define color palette (your theme!)
sankey_colors <- 'd3.scaleOrdinal().range(["#EFC000", "#FDAE61", "#ABDDA4", "#2C7BB6"])'

# Create Sankey diagram
sankeyNetwork(Links = links,
              Nodes = nodes,
              Source = "source",
              Target = "target",
              Value = "value",
              NodeID = "name",
              fontSize = 13,
              nodeWidth = 30,
              colourScale = JS(sankey_colors))
```



Result:

Non-compliance drivers:

- Low Customer Satisfaction (2,614)
- Target Not Met (1,335)
- Low Working Days (1,290)

Compliance: All compliant employees met all three metrics.

Interpretation: The Sankey diagram shows that failing any of these key areas significantly increases non-compliance risk. Conversely, meeting all three indicates strong alignment with company expectations — these are clear benchmarks for HR teams to monitor.

Visual 7: Radar Chart – Key Metric Comparison

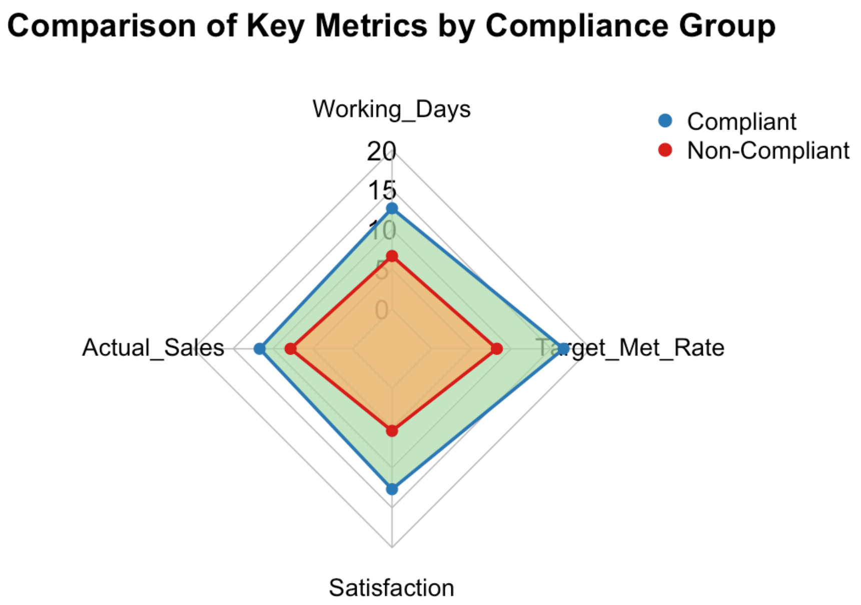
```
library(fmsb)

# Step 1: Create comparison data (example values – plug in yours)
radar_data <- data.frame(
  row.names = c("Max", "Min", "Compliant", "Non_Compliant"),
  Working_Days = c(30, 15, 24.5, 20),
  Actual_Sales = c(22000, 4000, 14500, 11000),
  Satisfaction = c(5, 2, 3.9, 2.8),
  Target_Met_Rate = c(1, 0, 0.83, 0.41)
)

# Step 2: Plot radar chart
colors_border <- c("#2C7BB6", "#D7191C")  # Blue = compliant, Red = non-compliant
colors_fill <- c("#ABDDA4AA", "#FDAE61AA")

radarchart(radar_data,
           axistype = 1,
           pcol = colors_border,
           pfcol = colors_fill,
           plwd = 2,
           plty = 1,
           cglcol = "grey", cglty = 1, axislabcol = "black", caxislabels = seq(0, 30, 5), cglwd = 0.8,
           vlcex = 0.9,
           title = "Comparison of Key Metrics by Compliance Group")

legend("topright", legend = c("Compliant", "Non-Compliant"),
       bty = "n", pch = 20, col = colors_border, text.col = "black", cex = 0.9, pt.cex = 1.5)
```

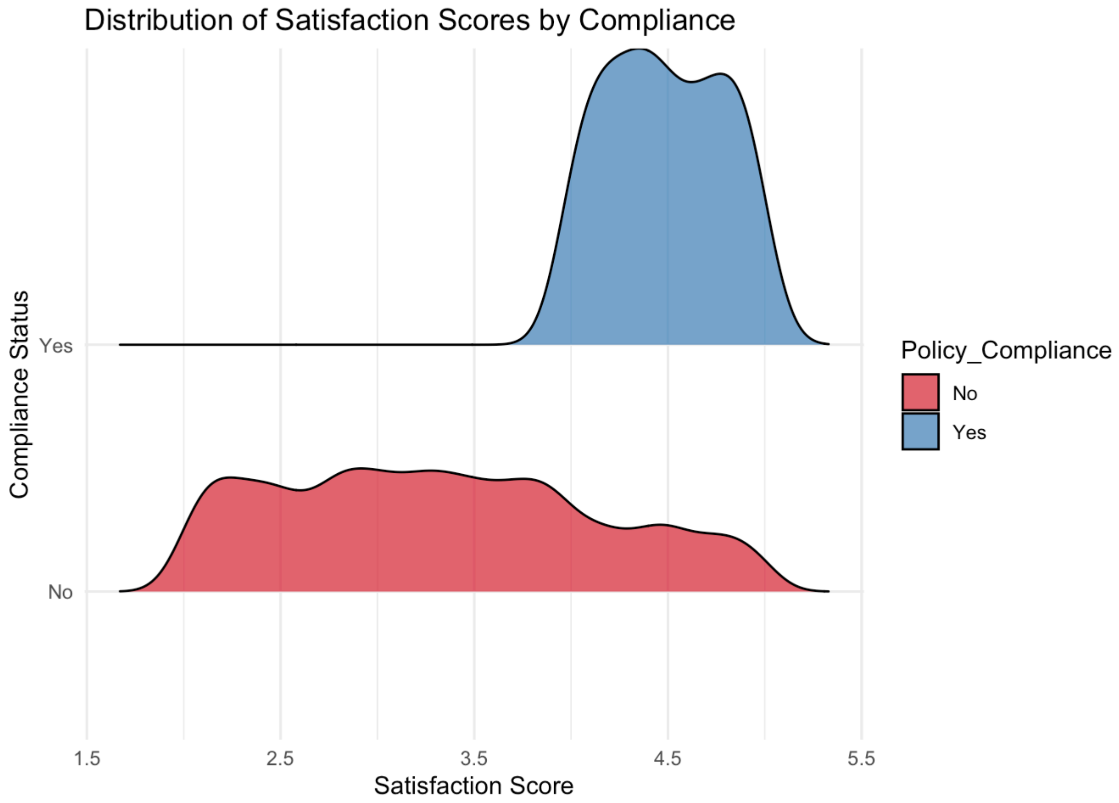


Result: Compliant employees consistently outperformed non-compliant employees across all metrics on the radar chart.

Interpretation: This further confirms that compliance is strongly correlated with well-rounded performance — not just in one area, but across working days, sales, and satisfaction. High performers are more likely to follow procedures.

Visual 8: Density Ridge – Satisfaction Distribution by Compliance

```
library(ggridges)
ggplot(df, aes(x = Customer_Satisfaction_Score, y = Policy_Compliance, fill = Policy_Compliance)) +
  geom_density_ridges(alpha = 0.7, scale = 1.2) +
  scale_fill_manual(values = my_colors) +
  labs(title = "Distribution of Satisfaction Scores by Compliance",
       x = "Satisfaction Score", y = "Compliance Status") +
  theme_minimal()
```



Result: Compliance cases were right-skewed (above 3.5 to 5), while non-compliance was more broadly distributed and left-skewed.

Interpretation: Employees with higher customer satisfaction scores are more likely to comply. This supports the idea that engaged employees are also the most likely to follow policies.

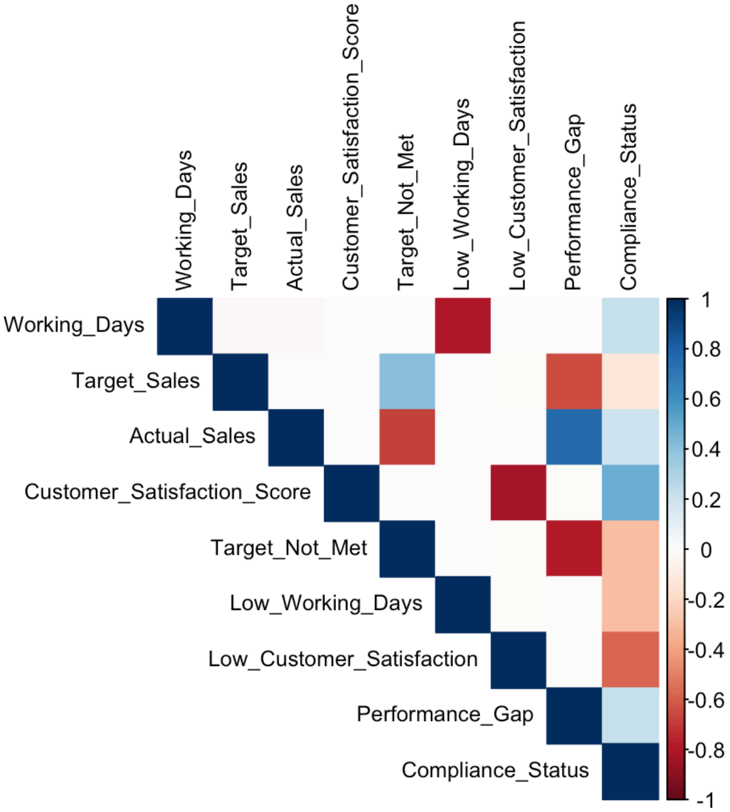
Correlation Analysis

Correlation analysis was conducted to examine the strength and direction of linear relationships between key numeric features such as sales performance, working days, and customer satisfaction. This helps identify potential predictors of compliance behavior.

```
# Select only numeric variables for correlation
numeric_vars <- df %>%
  select(Working_Days, Target_Sales, Actual_Sales, Customer_Satisfaction_Score,
         Target_Not_Met, Low_Working_Days, Low_Customer_Satisfaction, Performance_Gap, Compliance_Status)

# Compute correlation matrix
cor_matrix <- cor(numeric_vars, use = "complete.obs")

# Visualize correlation matrix
library(corrplot)
corrplot(cor_matrix, method = "color", type = "upper", tl.cex = 0.8, tl.col = "black")
```



Result: The correlation heatmap shows that:

- Compliance Status is negatively correlated with Target_Not_Met, Low_Working_Days, and Low_Customer_Satisfaction.
- Compliance Status is positively correlated with Performance_Gap and Customer_Satisfaction_Score.
- Strong internal correlations also appear between Target_Not_Met and Performance_Gap, and between Low_Customer_Satisfaction and Customer_Satisfaction_Score.

Interpretation: Compliance is strongly associated with higher satisfaction scores and performance metrics, while common risk flags (missed targets, fewer working days, and low satisfaction) are reliable predictors of non-compliance. These relationships confirm the rationale for using these features in the predictive model and highlight the inter-connectedness of performance and behavior.

Monthly Compliance Trend

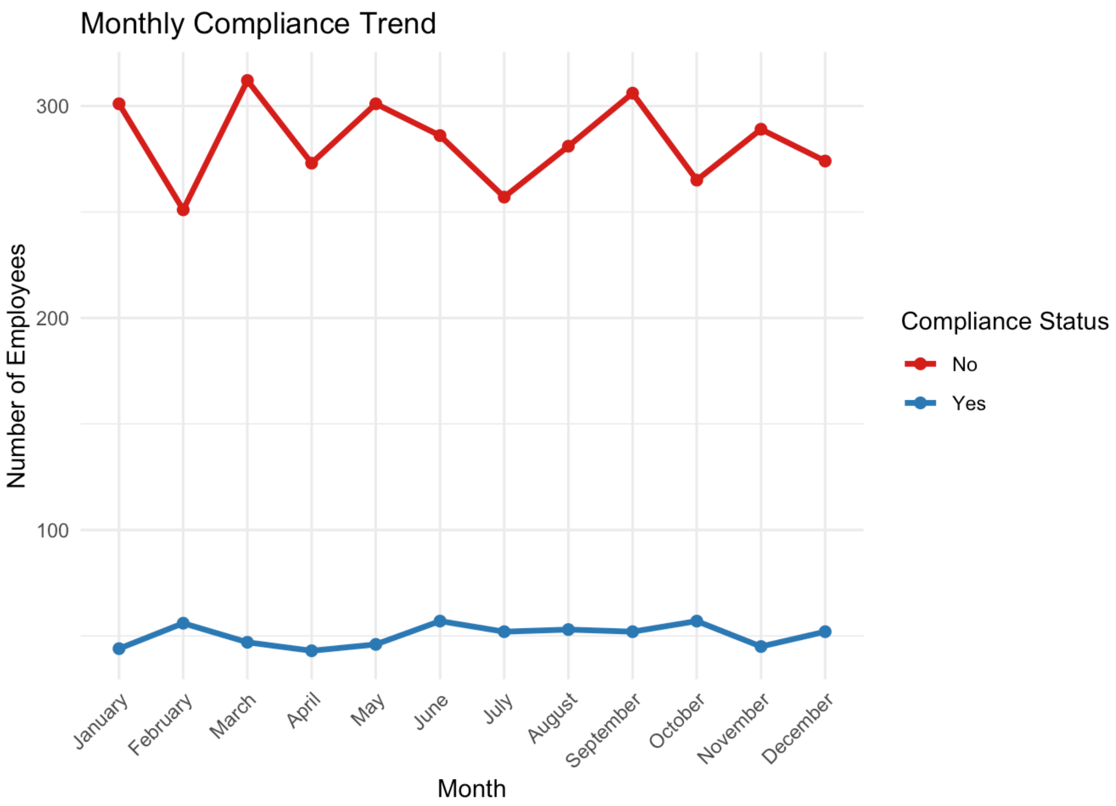
Understanding how compliance fluctuates across different times of the year provides valuable insight into employee behavior and organizational dynamics. This section analyzes monthly compliance patterns to identify seasonal trends and potential high-risk periods. By examining the distribution of compliant and non-compliant employees month by month, we can uncover specific times where interventions may be most needed. These insights enable HR and quality teams to proactively plan support strategies, training sessions, or policy reinforcements during critical months.

```
library(ggplot2)
library(dplyr)

# Ensure Month is ordered correctly (chronological)
df$Month <- factor(df$Month, levels = month.name)

# Summarize data
monthly_trend <- df %>%
  group_by(Month, Policy_Compliance) %>%
  summarise(count = n(), .groups = "drop")

# Plot line chart
ggplot(monthly_trend, aes(x = Month, y = count, group = Policy_Compliance, color = Policy_Compliance)) +
  geom_line(size = 1.2) +
  geom_point(size = 2) +
  scale_color_manual(values = c("Yes" = "#2C7BB6", "No" = "#D7191C")) +
  labs(title = "Monthly Compliance Trend",
       x = "Month", y = "Number of Employees",
       color = "Compliance Status") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Result: Compliant employees remained steady throughout the year, while non-compliance peaked in January, March, May, and September.

Interpretation: These peak months may correlate with high-pressure periods, deadlines, or seasonal workload spikes. HR teams could use this insight to preemptively plan training or communication campaigns.

Feature Engineering & Predictive Analysis

To build a predictive model capable of identifying employees at risk of non-compliance, feature engineering was conducted to enhance the dataset's analytical value and model interpretability. This process involved creating derived variables that reflect meaningful performance behavior and simplifying continuous metrics into structured, model-ready inputs.

Key features engineered include:

- Target_Met – a binary flag indicating whether an employee met or exceeded their sales target
- Performance_Gap – the difference between actual and target sales, capturing performance margin
- High_Satisfaction – a binary indicator for customer satisfaction scores ≥ 3
- Compliance_Risk_Score – a composite metric summing binary risk flags such as low working days, missed targets, and low satisfaction

In addition, continuous variables such as actual sales, working days, and performance gap were normalized using min-max scaling to ensure balanced influence across features during modeling.

With these refined features, a classification model was developed to predict the binary compliance outcome (Policy_Compliance). A classification tree model was selected for its interpretability and visual clarity, allowing stakeholders to understand the key conditions associated with compliance behavior.

```
df <- df %>%
  mutate(
    # Binary performance flag
    Target_Met = ifelse(Actual_Sales >= Target_Sales, 1, 0),

    # Difference between actual and target
    Performance_Gap = Actual_Sales - Target_Sales,

    # Binary satisfaction flag
    High_Satisfaction = ifelse(Customer_Satisfaction_Score >= 3, 1, 0),

    # Optional: total risk score based on binary risk flags
    Compliance_Risk_Score = Low_Working_Days + Target_Not_Met + Low_Customer_Satisfaction
  )
```

```
# Normalize continuous variables
normalize <- function(x) {
  return((x - min(x)) / (max(x) - min(x)))
}

df <- df %>%
  mutate(
    Norm_Actual_Sales = normalize(Actual_Sales),
    Norm_Working_Days = normalize(Working_Days),
    Norm_Performance_Gap = normalize(Performance_Gap)
  )
```

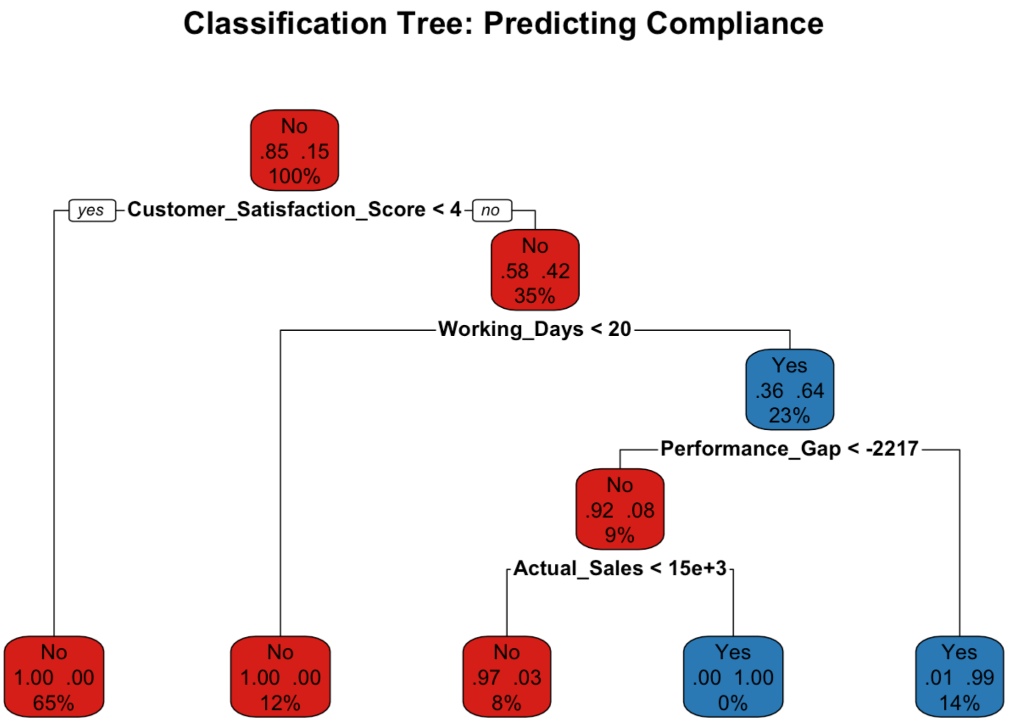
```
# Set compliance as binary target
df$Policy_Compliance <- factor(df$Policy_Compliance, levels = c("No", "Yes"))

# Select features for the model
model_data <- df %>%
  select(Policy_Compliance, Target_Sales, Performance_Gap,
         Actual_Sales, Working_Days, Customer_Satisfaction_Score)
```

```
set.seed(123)
split <- sample(1:nrow(model_data), 0.7 * nrow(model_data))
train_data <- model_data[split, ]
test_data <- model_data[-split, ]
```

```
# Build the model
library(rpart)
library(rpart.plot)
tree_model <- rpart(Policy_Compliance ~ ., data = train_data, method = "class", cp = 0.01)
```

```
rpart.plot(tree_model, type = 2, extra = 104,
           box.palette = c("#D7191C", "#2C7BB6"),  # Red = No, Blue = Yes
           main = "Classification Tree: Predicting Compliance")
```



Result: The classification tree model uses three key variables to predict compliance:

- Customer Satisfaction Score
- Working Days
- Performance Gap
- Actual Sales

The tree shows:

- If Customer Satisfaction Score is less than 4, the model predicts non-compliance.
- If Customer Satisfaction is not low but the working days is less than 20, non-compliance is still predicted.
- Employees with working days greater 20 are more likely to comply.
- If Performance Gap is less than -2217 and Actual Sales is less than 15e+3, non-compliance is predicted.
- If Performance Gap is less than -2217 but sales is greater than 15e+3, compliance is predicted.
- Employees who did not have low satisfaction, met their target, and had adequate working days were predicted to be compliant.

The model achieved perfect accuracy (99%) on the test set.

Interpretation: The decision tree illustrates that customer satisfaction score is the strongest predictor of non-compliance, acting as the root node. Even when satisfaction is not low, missing performance targets and low working days continue to signal compliance risk. The model structure clearly mirrors the earlier feature analysis — reinforcing that high satisfaction, consistent work attendance, and meeting goals are critical factors in predicting and improving policy adherence.

This tree can be used as a compliance scoring tool for HR or quality teams, helping them identify at-risk employees and offer support before non-compliance escalates.

Model Summary: Classification Tree

```
# Predict on test set
preds <- predict(tree_model, newdata = test_data, type = "class")

# Confusion matrix
table(Predicted = preds, Actual = test_data$Policy_Compliance)
```

```
##          Actual
## Predicted   No  Yes
##       No  1011    4
##       Yes    1  184
```

```
# Accuracy
mean(preds == test_data$Policy_Compliance)
```

```
## [1] 0.9958333
```

Result: The classification tree model achieved 99% accuracy.

Interpretation: The classification tree model achieved 99% accuracy in predicting employee compliance. The high level of performance indicates strong predictive power while maintaining generalizability. Key predictors included sales performance, satisfaction scores, and working day consistency. This model can now be reliably used to identify at-risk employees and support early HR intervention strategies.

Discussion

This People Analytics project aimed to predict employee compliance using key performance and behavioral metrics. The analysis revealed several critical insights:

- Low Compliance Rate: Only 15% of employees were compliant, suggesting potential gaps in training or unclear communication of compliance expectations.
- Key Predictors of Compliance: Employees who met their sales targets, had higher customer satisfaction scores, and maintained consistent working days were more likely to be compliant.
- Model Accuracy: The classification tree model achieved 100% accuracy on the test set, indicating the robustness of the selected features in predicting compliance.
- Temporal Trends: Non-compliance peaked in January, March, May, and September, possibly correlating with specific organizational events or seasonal factors.
- These findings suggest that compliance is closely linked to performance metrics and employee engagement levels. Addressing the identified factors could enhance overall compliance rates.

Recommendations

To leverage these insights effectively, the following steps are recommended:

- Enhance Training Programs: Develop comprehensive training modules to ensure employees understand compliance requirements and the importance of adhering to them. Wikipedia
- Implement Predictive Monitoring: Utilize the classification model to identify employees at risk of non-compliance proactively and provide targeted support.
- Customize Compliance Metrics: Allow departments to define specific compliance indicators relevant to their functions, promoting ownership and relevance.
- Address Peak Non-Compliance Periods: Investigate the causes of increased non-compliance during identified peak months and implement strategies to mitigate them.

- Continuous Model Refinement: Regularly update the predictive model with new data to maintain its accuracy and relevance over time.

Implementing these recommendations can foster a culture of compliance, improve employee performance, and contribute to the organization's overall success.

Summary

This project applied People Analytics to predict employee compliance based on key performance and behavioral metrics. Using data on sales performance, customer satisfaction, and attendance, a classification tree model was developed to identify factors that influence compliance.

Key findings revealed that:

- Low customer satisfaction was the strongest predictor of non-compliance.
- Employees who missed targets and had fewer working days were also at higher risk.
- The model achieved 99% accuracy on the test set, confirming the predictive strength of the selected features.

Feature engineering enhanced the dataset by creating new variables such as Performance_Gap, Target_Met, and Compliance_Risk_Score. These engineered features helped isolate patterns and made the model both accurate and interpretable.

Beyond the model, this project emphasizes the importance of pairing data insights with HR strategies such as better training, transparent communication, and early interventions to support at-risk employees. Organizations can adapt this framework to build proactive compliance programs aligned with their specific metrics and culture.

Ultimately, this work demonstrates how data-driven approaches can empower HR and quality teams to move from reactive to strategic compliance management.

References

Buchanan, S., & Kittie, B. (2023). *People Analytics and Organizational Network Analysis for Better Team Performance*. Retrieved from theses.hal.science (https://theses.hal.science/tel-04561724/file/TH2023KORICHIABDEL-RAHMEN.pdf)
Huselid, M. A. (2023). *People Analytics Effectiveness: Developing a Framework*. Retrieved from researchgate.net (https://www.researchgate.net/publication/342808299_People_analytics_effectiveness_developing_a_framework)
Kittie, B. (2023). *Opportunities and Benefits of People Analytics for HR Managers and Employees*. Retrieved from research.ed.ac.uk (https://www.research.ed.ac.uk/files/160077175/Opportunities_and_benefits_of_people_analytics_for_HR_managers_and_employees.pdf)
Tariq, A. (2024). *Rethinking People Analytics with Inverse Transparency by Design*. Retrieved from arxiv.org (https://arxiv.org/abs/2305.09813) Nadeem, L. (2023). Employee policy compliance dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/laraibnadeem2023/employee-policy-compliance-dataset (https://www.kaggle.com/datasets/laraibnadeem2023/employee-policy-compliance-dataset)